SceneBind: Binding What and Where Across Vision, Audio, and Language

Mingfei Chen[1], **Zijun Cui**[1,2]*, **Ruoke Zhang**[1], **Hyeonggon Ryu**[3], **Eli Shlizerman**[1]†
[1]University of Washington [2]University of Texas at Dallas
[3]Hankuk University of Foreign Studies

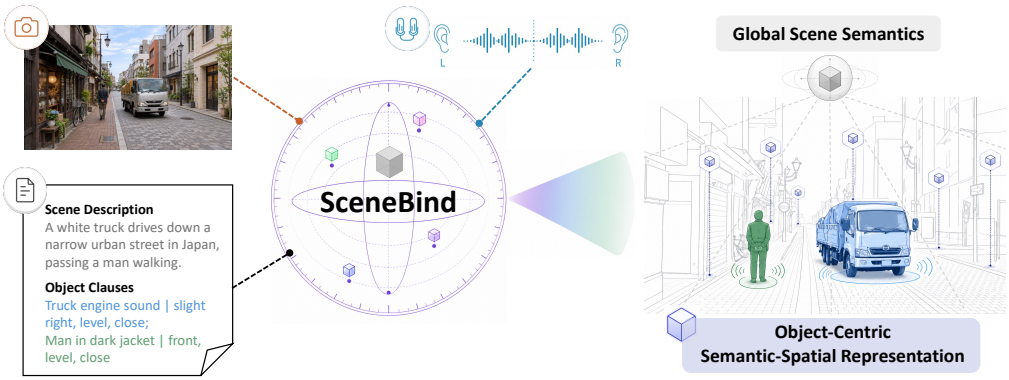

Figure 1: SceneBind models scenes by jointly capturing *what* (semantics) and *where* (spatial) across vision, audio, and language. SceneBind maps each modality into a shared global embedding with object-centric slots, enabling semantic–spatial representation of multimodal scenes.

Abstract

We present **SceneBind**, an omni-modal representation of realistic scenes with joint semantic and 3D spatial understanding across vision, audio, and language. Existing omni-modal encoders excel at instance-level semantics (i.e., *what* is present), but often lack explicit spatial structure (i.e., *where* it is). SceneBind addresses this gap by representing each scene as a semantic-spatial entity, combining a global semantic embedding with object-centric semantic-spatial slots. This representation explicitly captures object-level semantics, spatial attributes, and uncertainty. We further propose **SceneBind Matching**, a semantic–spatial matching scheme that integrates global scene similarity with object alignment, supporting cross-modal scene retrieval and object grounding. To train and evaluate SceneBind, we curate a novel real-world binaural audio-visual dataset with structured semantic and spatial annotations, and propose a training protocol for aligning semantic and spatial signals across modalities. SceneBind is compatible with large-scale pretrained semantic encoders, adds lightweight spatial modeling with only a few additional tokens. It achieves state-of-the-art scene and spatial retrieval while enabling strong zero-shot transfer to downstream tasks such as audio-visual localization.

*Led the data curation work while at the University of Washington
†Corresponding author: shlizee@uw.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Understanding real world environments emphasizes semantic and spatial perception which captures both *what* is present and *where* it is in 3D space. This capability is essential for multimodal spatial intelligence [1, 2], enabling spatial reasoning and interaction across vision, audio, and language, and supporting applications in embodied AI, robotics, and spatially grounded multimodal world models.

Recent advances in large-scale multimodal learning have yielded powerful representations [3, 4, 5, 6, 7, 8] that align semantic concepts of a scene across modalities such as images and audio. While highly effective for semantic understanding, these models provide limited explicit modeling of spatial structure. In this work, we aim to learn a representation that unifies semantics and spatial structure for real-world omni-modal scenes, where each modality is encoded into a shared semantic–spatial space.

Achieving this goal presents several challenges. First, semantic embeddings are typically global, whereas spatial information is inherently object-centric. In complex scenes, a single embedding is often insufficient to represent multiple objects with distinct spatial configurations. Second, cross-modal spatial learning is data-limited. Real-world datasets have advanced vision-audio-language alignment, but explicit spatial annotations remain limited, and spatial audio requires careful curation to disentangle sound sources and label directional properties [2]. Synthetic acoustic simulators [9] offer useful partial solutions, but still face realism gaps and misalignment with real-world visual data.

To address these challenges, we propose **SceneBind**, a framework for learning semantic–spatial representations of omni-modal scenes. As shown in Figure 1, SceneBind represents a scene through a global semantic embedding and object-centric semantic-spatial representations. It captures both holistic semantic context and the location of sources in a camera-relative space. This representation can provide a lightweight spatial prior for downstream tasks, e.g., localization. To support this formulation, we introduce a data curation pipeline based on binaural videos, constructing a novel omni-modal dataset and benchmark with structured spatial annotations for real-world scenes.

Building on SceneBind representation and dataset, we develop a training strategy for learning object-centric representations with varying numbers of objects in a scene. We use bipartite matching to assign predicted object slots to ground-truth semantic-spatial object descriptions, and combine global semantic alignment, object grounding, and object-centric contrastive objectives to learn discriminative, spatially grounded representations. We further introduce **SceneBind Matching** for cross-modal scene comparison, enabling retrieval, grounding, and zero-shot tasks such as audio-visual localization.

Our contributions are: (1) We introduce **SceneBind** for learning omni-modal scene representations that capture *what* is present (semantics) and *where* it is (spatial location) through global and object-centric joint representations. (2) We curate a novel real-world binaural audio-visual dataset with structured spatial annotations to support training and benchmarking semantic-spatial scene representations. (3) We propose **SceneBind Matching**, combining global and object-level alignment for cross-scene comparison, retrieval, and grounding. (4) SceneBind integrates with pretrained encoders with minimal overhead, achieving state-of-the-art performance and zero-shot generalization.

2 Related Works

2.1 Multimodal Pretraining

Large-scale contrastive pretraining has enabled strong cross-modal alignment in a shared embedding space. Image-text models such as CLIP [3] and SigLIP [5, 4] learn transferable visual representations, while audio-text models such as CLAP [6, 7, 10] extend this paradigm to sound. More general encoders further bind multiple modalities: AudioCLIP [11] aligns audio and vision through text, and ImageBind [8] uses image-based alignment to unify image, text, audio, and other signals, supporting retrieval and zero-shot transfer. These encoders are effective semantic foundations for multimodal systems [12, 13, 14], and focus on emphasizing semantics (*what* is present) and may omit location (*where* entities are). Spatial cues are often weakly preserved: rotations and flips may be treated as semantic equivalents, and audio is frequently downmixed to monaural signals. SceneBind addresses this gap by preserving spatial factors while remaining compatible with alignment-based pretraining.

2.2 Spatially Grounded Multimodal Representations

Multimodal spatial reasoning requires estimating where entities are and how they relate in 3D, including direction, distance, and relative geometry [1, 2]. Vision-centric approaches model 3D structure with point clouds [15], graphs [16, 17, 18], voxel grids [19], maps [20, 1], and neural fields [21, 22]. Object-centric methods, including detection and grounding [23, 24, 25, 26, 27] and multi-modal scene graphs [28, 29], further capture fine-grained scene structure, but are often local, foreground-focused, or not designed as unified cross-modal representations.

Audio is an important modality for identifying and spatially localizing objects and events that produce sound. However, many audiovisual spatial grounding use monaural audio and rely on visual cues to localize sound [30, 31, 32, 33, 34], making spatial inference challenging under occlusion, out-of-view sources, or dynamic scenes. Spatial audio provides direct physical cues, such as inter-channel phase and level differences, for estimating direction and distance [35, 36, 37, 10, 38, 39, 40]. When spatial audio was taken into account in prior works, it was often from synthetic data or constrained sensors, limiting generalization to noisy real-world environments. In contrast, SceneBind learns unified semantic-spatial representations from real-world audio, vision, and language, combining global semantics with semantic-spatial slots to capture holistic scene context and object-level alignment.

3 Method

3.1 Problem Setup and System Overview

We define a "*scene*" as a semantic-spatial entity that captures *what* is present and *where* sources lie in camera-relative space. A scene can be observed through multiple modalities, including image I, audio A, and text T. The text modality consists of a global scene description and a set of object-level clauses, each paired with spatial attributes $\mathbf{r}_k = (\theta_k, \phi_k, d_k)$ denoting coarse camera-relative azimuth, elevation, and distance categories. SceneBind maps each modality into a shared semantic–spatial representation $\mathcal{X}$ consisting of a global embedding and K object slots.

$$\mathcal{X} = \left(\mathbf{s}_{\text{global}}, \{(\mathbf{s}_k, \mathbf{r}_k, c_k)\}_{k=1}^{K}\right), \tag{1}$$

where $\mathbf{s}_{\text{global}} \in \mathbb{R}^d$ encodes scene-level semantics, and each object slot consists of a semantic embedding $\mathbf{s}_k \in \mathbb{R}^d$, spatial attributes $\mathbf{r}_k$, and a confidence score c_k.

As illustrated in Figure 2, SceneBind encodes each modality into a shared semantic-spatial representation $\mathcal{X}$ composed of a global embedding and object-centric semantic-spatial slots (Sec. 3.2). It is trained with global alignment and bipartite matching supervision (Sec. 3.3) to align global-level semantics across modalities and ground predicted slots to ground-truth (GT) text clauses. For inference, SceneBind Matching compares scene representations by combining global semantic similarity with object-centric slot alignment over $(\mathbf{s}_k, \mathbf{r}_k, c_k)$ (Sec. 3.4).

3.2 SceneBind Encoding Model

Given an input sample from modality $m \in \{A, V\}$, SceneBind uses modality-specific encoders and decoders to extract semantic–spatial context tokens, obtain global and object semantic states, and decode object spatial attributes. Together, these modules produce the scene representation $\mathcal{X}^m$.

Semantic-Spatial Context Tokens. For vision, we use the frozen visual tower of a vision-text pretrained encoder SigLIP2 [5] to produce patch tokens $\mathbf{F}^V \in \mathbb{R}^{L_V \times d}$ in the vision-text space, where the patch sequence preserves image layout and provides both semantic and spatial evidence. For audio, a frozen M2D-CLAP [7] audio tower processes the mono mixture to semantic patch tokens in the audio-text space. We train a lightweight alignment module to further map these tokens into the vision-text embedding space, yielding $\mathbf{F}^A_{\text{sem}} \in \mathbb{R}^{L_A \times d}$. To capture spatial cues, we convert the binaural waveform into a four-channel time-frequency representation, consisting of left/right magnitude and phase, and feed it to a convolutional spatial encoder. The resulting spatial features are projected to dimension d, interpolated to length L_A, and added to $\mathbf{F}^A_{\text{sem}}$ to form the audio context $\mathbf{F}^A \in \mathbb{R}^{L_A \times d}$.

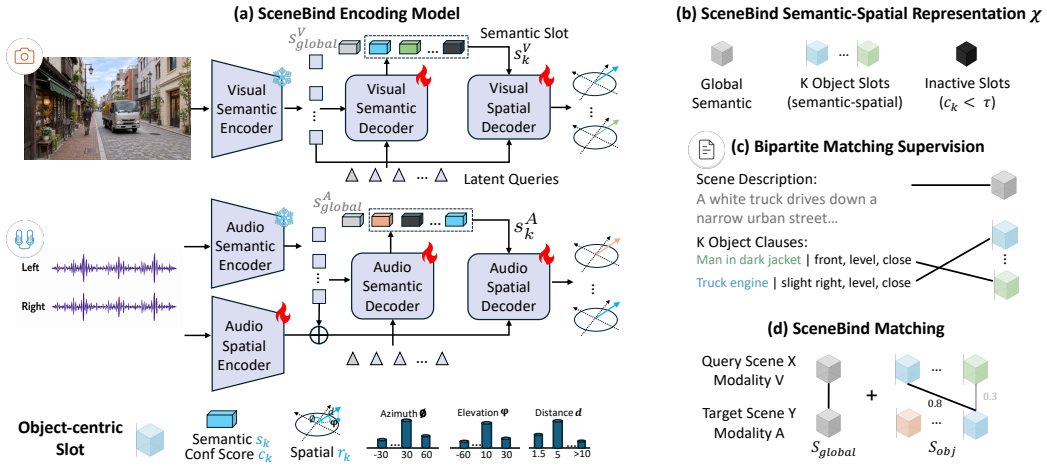

Figure 2: **SceneBind overview.** (a) **SceneBind Encoding Model** maps multimodal inputs to a global embedding $\mathbf{s}^{\text{global}}$ and object-centric slots $(\mathbf{s}_k, \mathbf{r}_k, c_k)$. (b) **SceneBind Representation** $\mathcal{X}$ represents each scene with a global embedding and K object-centric semantic-spatial slots. (c) Bipartite matching aligns slots with GT object clauses for supervision. (d) **SceneBind Matching** combines global similarity S_{global} with object slot alignment score S_{obj} for cross-modal scene comparison.

Semantic Decoding. SceneBind uses a cross-attention semantic decoder D^m_{sem} to extract global and object semantics from visual tokens $\mathbf{F}^V$ for vision and audio semantic tokens $\mathbf{F}^A_{\text{sem}}$ for audio. A learned global query aggregates holistic scene content, analogous to pooling in pretrained encoders [5, 7], and outputs a global embedding $\mathbf{s}^m_{\text{global}}$. Distinctively, SceneBind introduces K learned latent object queries that act as semantic anchors, decoding object semantic states $\{\mathbf{h}^m_k\}_{k=1}^K$ from context tokens. Each object state is decoded to a semantic slot $\mathbf{s}^m_k$ and activity confidence c^m_k

$$\mathbf{s}^m_k = g^m_{\text{sem}}(\mathbf{h}^m_k), \qquad c^m_k = \sigma(g^m_{\text{conf}}(\mathbf{h}^m_k)). \tag{2}$$

As a result, audio and vision modalities use separate learned queries, while the outputs of these queries are shared between modalities in the semantic-spatial space.

Spatial Decoding. The spatial decoder D^m_{spa} uses each object semantic slot $\mathbf{s}^m_k$ from D^m_{sem} as a query to decode spatial attributes $\hat{\mathbf{r}}^m_k = (\hat{\theta}^m_k, \hat{\phi}^m_k, \hat{d}^m_k)$ from semantic-spatial context tokens. It cross-attends to visual patch tokens $\mathbf{F}^V$ for vision and to the fused audio context $\mathbf{F}^A$ for audio, where $\mathbf{F}^A$ combines semantic audio tokens with binaural spatial cues. The decoder output $\mathbf{z}^m_k$ is mapped by modality specific projection heads into spatial embeddings $\mathbf{u}^{m,\theta}_k$, $\mathbf{u}^{m,\phi}_k$, and $\mathbf{u}^{m,d}_k$, which are passed to shared azimuth, elevation, and distance classifiers. The final prediction $\hat{\mathbf{r}}^m_k$ is obtained by taking the argmax over logits of each classifier.

For text, the frozen text encoder [5] produces a global scene embedding from the scene description and object-centric semantic embeddings from individual clauses. Each annotation provides a semantic clause for text embedding and additional structured fields for spatial category targets.

3.3 Bipartite Matching Supervision

This section describes how we supervise the SceneBind representation, which consists of a global semantic embedding and object semantic-spatial slots. For global semantic supervision, we apply a symmetric InfoNCE [41] loss $\mathcal{L}_{\text{global}}$ to paired samples of audio-visual, audio-text and visual-text, using scene descriptions as text targets. This follows standard contrastive alignment, pulling matched samples together and pushing unmatched samples apart in the shared space.

Object-centric supervision is challenging because predicted slots are unordered. In each scene, there is a variable number of ground-truth object clauses. In SceneBind, a fixed set of K slots is predicted. As shown in Figure 2 (c), we use bipartite matching [42] to assign slots to ground-truth objects before applying object-level losses. Each ground-truth object j includes a semantic embedding $\mathbf{s}^*_j$

and spatial attribute $\mathbf{r}_j^* = (\theta_j^*, \phi_j^*, d_j^*)$. Thus for predicted slot i, the matching score is

$$a_{ij} = \big[(1-\lambda)\cos(\mathbf{s}_i, \mathbf{s}_j^*) + \lambda\,\rho_{ij}\big](\alpha + \beta c_i). \tag{3}$$

Here ρ_{ij} denotes the spatial matching score, computed as the average softmax probability assigned by slot i to the ground-truth azimuth, elevation, and distance bins of object j. The one-to-one assignment that maximizes the total matching score provides the slot supervision.

Object Grounding Supervision. The object grounding loss $\mathcal{L}_{\text{obj}}$ grounds each matched slot to its paired object clause in both semantics and space. We align the slot embedding to the clause embedding with cosine distance, and supervise azimuth, elevation, and distance using Gaussian-smoothed cross-entropy over discretized bins, so nearby spatial bins incur smaller penalties than distant ones, preserving local spatial continuity. Slot confidence is trained with binary cross-entropy, treating matched slots as positives and unmatched slots as negatives.

Object-Centric Contrastive Learning. Beyond grounding, the matched assignments define cross-modal object correspondences. This contrastive supervision encourages slots that refer to the same object to align across modalities, while keeping objects from different instances separable. We use an intra-scene loss $\mathcal{L}_{\text{inst}}$ and a cross-scene loss $\mathcal{L}_{\text{batch}}$. The intra-scene loss $\mathcal{L}_{\text{inst}}$ contrasts slots matched to the same ground-truth object across modalities as positives, and other slots within the same scene as negatives, encouraging separation of co-occurring objects. The cross-scene loss $\mathcal{L}_{\text{batch}}$ contrasts matched objects across the minibatch, further improving object discriminability. For audio-visual pairs, we apply cross-scene contrastive learning to both semantic and spatial embeddings

$$\mathcal{L}_{\text{ctr}}(U, V) = -\log \frac{\exp(\text{sim}(u_i, v_i)/\tau)}{\sum_j \exp(\text{sim}(u_i, v_j)/\tau)}, \tag{4}$$

$$\mathcal{L}_{\text{batch}}^{A,V} = \mathcal{L}_{\text{ctr}}(\mathbf{s}^A, \mathbf{s}^V) + \sum_{q\in\{\theta,\phi,d\}} \gamma_q\,\mathcal{L}_{\text{ctr}}(\mathbf{u}^{A,q}, \mathbf{u}^{V,q}), \tag{5}$$

where γ_q scales the contribution of each spatial axis, and matched audio-visual slots assigned to the same ground-truth clause are positives. For audio-text and visual-text pairs, $\mathcal{L}_{\text{batch}}$ uses only semantic embeddings because text has no learned spatial-head features.

Overall Training Objective. SceneBind is trained with a combination of global-level alignment ($\mathcal{L}_{\text{global}}$), object grounding ($\mathcal{L}_{\text{obj}}$), and object-centric contrastive objectives ($\mathcal{L}_{\text{inst}}, \mathcal{L}_{\text{batch}}$).

We train in two stages. The first stage uses mixed supervision from AudioCaps [43], MS-COCO [44], and our curated Binaural dataset with aligned spatial labels, combining broad semantic alignment with spatially grounded object-centric learning. The second stage further fine-tunes on the Binaural dataset only, specializing SceneBind for cross-modal semantic-spatial alignment at both global and object levels. Full training details are provided in the appendix.

3.4 SceneBind Matching

At inference time, we propose a unified matching scheme (Figure 2 (d)) that compares scenes using both global semantic alignment and object-centric semantic-spatial correspondence. Given a query scene $\mathcal{X}^q$ and a candidate scene $\mathcal{X}^c$, the global score is the cosine similarity between their scene-level embeddings, denoted as S_{global}. For object-centric matching, we first keep active slots according to their confidence scores. Each query slot retrieves its best-matching candidate slot under a semantic-spatial score weighted by both slot confidences

$$S_{\text{obj}} = \frac{\sum_i w_{ij_i^*}\,\cos(\mathbf{s}_i^q, \mathbf{s}_{j_i^*}^c)\,\rho_{ij_i^*}}{\sum_i w_{ij_i^*} + \epsilon}, \quad j_i^* = \arg\max_j w_{ij}\cos(\mathbf{s}_i^q, \mathbf{s}_j^c)\rho_{ij}, \quad w_{ij} = \sqrt{c_i^q c_j^c}. \tag{6}$$

Here ρ_{ij} measures spatial similarity between predicted azimuth, elevation, and distance distributions, and w_{ij} is the geometric-mean confidence, which gives higher weight to reliable slot pairs. Each query slot selects its best candidate independently. Finally, SceneBind combines global and object scores after query wise normalization: $S(\mathcal{X}^q, \mathcal{X}^c) = \text{norm}(S_{\text{global}}) + \lambda_{\text{match}}\,\text{norm}(S_{\text{obj}})$, where λ_{match} balances scene-level semantic alignment and object-level semantic-spatial matching.

4 Experiments

4.1 Spatially Grounded In-the-Wild Data Curation

We curate spatially grounded omni-modal data from two sources. The primary *Binaural* corpus is collected from 21,927 in-the-wild videos with binaural audio, providing real spatial audio cues. Our multi-stage pipeline filters source videos, annotates candidate events, verifies cross modal consistency, balances spatial and semantic coverage, and applies human review for evaluation. We use Gemini [45] to propose short events from video clips, linking audible events to visible entities and estimating camera-coordinate spatial labels from visual evidence. The proposals are then verified with pretrained audio and visual encoders [8, 6, 7] and balanced to reduce spatial and semantic redundancy. Each retained clip contains a two-second audio-image pair, a scene description, and object clauses with instance semantic description and spatial labels (azimuth, elevation, and distance). The final training split contains 38,430 clips with 207,836 objects, and the human-verified Binaural benchmark contains 1,066 clips. Four reviewers corrected spatial categories using temporal context and binaural playback, removing ambiguous object. Additional verification and agreement details can be found in Appendix Sec. B.3 We also build a *Sphere360* benchmark from 360° ambisonic videos [46] for zero-shot evaluation under full-sphere viewpoint variation. More details are provided in the appendix.

4.2 Tasks and Evaluation

We evaluate SceneBind to answer: (1) How does SceneBind improve over global semantics-only representations across retrieval and object grounding tasks? (2) How do training and inference designs affect semantic-spatial alignment? (3) How does SceneBind enable zero-shot generalization?

For evaluation of SceneBind's Binaural scene retrieval, spatial retrieval, object grounding, and Sphere360 retrieval, we report the mean $\pm$ sample standard deviation over three random seeds.

Cross-modal Scene Retrieval. We evaluate retrieval across audio (A), visual (V), and text (T) modalities (Table 1), where each scene includes binaural audio, an image, and a text description (global caption + object clauses). Given a query in one modality (or a fused pair), the task is to retrieve the corresponding scene in another modality. We report bidirectional Recall@1 (R@1), i.e., the fraction of queries whose correct match is ranked first, for $A \leftrightarrow V$, $A \leftrightarrow T$, and $V \leftrightarrow T$, with modality-specific retrieval pools detailed in Appendix Sec. B.4. We further evaluate zero-shot generalization on Sphere360 (Table 7), where each text query uses a single dominant object clause with the global caption and candidates are spatially augmented views of the same scene, requiring fine-grained spatial discrimination.

Spatial Retrieval. We evaluate spatial retrieval using object text clauses with spatial attributes (Table 1). Candidates are ranked by the maximum spatial similarity over object-centric predictions within each sample. We report spatial mAP, which measures how highly candidates with the correct spatial label are ranked, and spatial R@1, computed within the same semantic cluster but different spatial configurations to test spatial discrimination.

Object Grounding. Object grounding evaluates the semantic identity and coarse camera-relative spatial categories of a queried object. Given an object query, we encode its semantic description (without spatial attributes) and match it to predicted object slots based on semantic similarity and confidence (Table 3). The best-matching slot is selected for evaluation. We report recall as the fraction of queries with semantic similarity > 0.5, and accuracy as those additionally with all spatial labels (azimuth, elevation, distance) correct. We also report per-attribute spatial accuracy.

Zero-shot Downstream Tasks. We evaluate zero-shot transfer to multimodal tasks requiring joint spatial and semantic reasoning via egocentric audio-visual localization [47] (Table 8). Given stereo audio and an image, the goal is to predict the sounding region in the image. We directly apply SceneBind without task-specific training. We report *cIoU* at multiple thresholds (e.g., 0.2, 0.3, 0.4), measuring overlap between predicted and GT, and *AUC* summarizing performance across thresholds.

4.3 Baselines

We compare with representative multimodal encoders, including AudioCLIP [11], CLAP variants (LAION-CLAP [6], M2D-CLAP [7]), SpatialCLAP [10], and ImageBind [8]. We pair CLAP-based audio encoders with SigLIP2 [5], while AudioCLIP and ImageBind use native encoders. For zero-shot

Table 1: **Cross-modal scene retrieval (R@1) and object-level spatial retrieval** on unseen Binaural test set. SceneBind outperforms all baselines, with largest gains on text-conditioned retrieval (especially V↔T) and spatial metrics, demonstrating effective semantic–spatial representation for scenes.

Encoder		Scene Retrieval (R@1 ↑)				Spatial Retrieval (↑)		
Audio–Text	Visual–Text	A↔V	A↔T	V↔T	Avg	R@1	mAP	Avg
AudioCLIP [11]	AudioCLIP [11]	0.3	0.3	20.6	7.1	9.3	28.1	18.7
SpatialCLAP [10]	SigLIP2 [5]	0.4	1.9	<u>51.2</u>	17.8	10.1	28.0	19.0
LAION-CLAP [6]	SigLIP2 [5]	0.5	5.9	<u>51.2</u>	19.2	10.1	28.7	19.4
M2D-CLAP [7]	SigLIP2 [5]	0.3	9.0	<u>51.2</u>	20.2	<u>11.4</u>	<u>29.6</u>	<u>20.5</u>
ImageBind [8]	ImageBind [8]	7.8	7.6	48.0	21.2	10.6	28.7	19.6
M2D-CLAP* [7]	SigLIP2* [5]	<u>19.3</u>	<u>11.2</u>	42.8	<u>24.4</u>	10.1	28.9	19.5
ImageBind* [8]	ImageBind* [8]	<u>19.3</u>	10.9	35.4	21.9	10.4	29.4	19.9
SpatialCLAP* [10]	SigLIP2* [5]	16.1	9.3	44.2	23.2	10.6	29.2	19.9
SceneBind (Ours)	**SceneBind (Ours)**	**21.3** ± 0.6	**17.6** ± 0.6	**65.5** ± 0.7	**34.8** ± 0.4	**27.9** ± 1.3	**47.1** ± 1.0	**37.5** ± 1.2

evaluation, we use native text encoders. Baselines without joint audio–visual alignment, including CLAP-based encoders paired with SigLIP2, use chaining over the shared benchmark text inventory. Its native audio–text and SigLIP2 text embeddings, T_{AT} and T_{VT}, give $S_{\mathrm{AV}} = (AT_{\mathrm{AT}}^{\top})(T_{\mathrm{VT}}V^{\top})$, summing similarity products over all intermediate texts. As these models produce a single embedding per sample, we treat them as a single object-centric prediction (confidence 1.0, no spatial attributes), match each clause independently, and aggregate similarities for retrieval. For further fair comparison, we fine-tune a selected set of strong and representative baselines, including ImageBind*, M2D-CLAP* paired with SigLIP2*, and SpatialCLAP* paired with SigLIP2* (which supports spatial audio input), under the same data and training setup as SceneBind by learning projection heads into a shared space aligned with frozen SigLIP2 text features, using global and multi-positive (object clause) contrastive losses. Additional details are provided in the appendix.

4.4 Main Results

Cross-modal Scene Retrieval. SceneBind achieves 34.8 ± 0.4 overall scene retrieval, exceeding the best fine-tuned baseline by 10.4 points (Table 1). V↔T improves by 28% over pretrained and 48% over fine-tuned baselines; A↔T improves by 57%. The largest gains are on text-relevant retrieval—especially V↔T. This highlights that joint semantic-spatial modeling with object-centric matching of SceneBind improves scene retrieval beyond global semantic encoding..

On Sphere360 (Table 7), SceneBind achieves an overall score of 27.3 ± 1.7, improving over the best fine-tuned baseline, ImageBind*, by 8.5 percentage points (+45%). Gains concentrate in text-related retrieval: V↔T improves by 8.1 points and A↔T by 17.5 points, while A↔V matches ImageBind* at 23.3. These results support strong zero-shot spatial discrimination under semantic ambiguity.

Spatial Retrieval. SceneBind improves spatial R@1 and mAP over the best reported baselines by 16.5 and 17.5 points (+145% and +59%), respectively (Table 1). Notably, fine-tuning existing encoders yields minimal gains, since they mostly rely on global semantics and lack explicit spatial grounding. In contrast, SceneBind's object-centric spatial representation and semantic–spatial matching enable accurate retrieval within semantically identical but spatially varying candidates.

Object Grounding. SceneBind enables text-guided object grounding via its object-centric semantic-spatial representation. Audio grounding favors azimuth and elevation, while vision favors left/right discrimination and distance (Table 3). The mono control shows the strongest binaural benefit in horizontal discrimination and gains on non-level elevation bins (Appendix Sec. E). Audio provides stronger elevation recovery, whereas vision performs better on left/right azimuth and distance estimation. Together, these results demonstrate that SceneBind learns meaningful associations between object text semantics and spatial attributes across both modalities.

4.5 Ablations

Training Objectives Ablations. Removing object-level semantic supervision (no $\mathcal{L}_{\mathrm{sem}}$) degrades grounding (19.6→8.8) and spatial retrieval (38.4→35.2), as object representations are no longer aligned with text queries, while scene retrieval slightly increases (34.7→36.0) due to global bias.

Table 2: **Binaural cues for spatial retrieval and grounding.** Sphere360 hard scene retrieval distinguishes semantically similar scenes with different spatial layouts. Binaural evaluates spatially conditioned audio retrieval and left/right (L/R) grounding.

Input	Hard scene retrieval A↔V R@1 ↑	Hard scene retrieval A↔T R@1 ↑	Audio spatial retrieval R@1 ↑	Audio spatial retrieval mAP ↑	Audio grounding L/R accuracy ↑
Binaural	25.7	30.1	23.2	43.9	62.7
Mono	25.2	24.8	17.7	38.2	50.8
Δ	-0.5	-5.3	-5.5	-5.7	-11.9

Removing intra scene contrastive loss ($\mathcal{L}_{\text{inst}}$) leads to overall degradation (Avg 30.9→27.2), with notable drops in spatial (38.4→31.1) and grounding (19.6→16.1), indicating weaker object discrimination. Removing batch level contrastive loss ($\mathcal{L}_{\text{batch}}$) primarily affects spatial retrieval (38.4→36.5), as it encodes spatial discrimination across scenes, while intra scene alignment still preserves semantic matching. Combining all losses yields the best overall performance.

Data Recipe with Two-Stage Training. We train SceneBind with multiple source of data, including our novel Binaural dataset (spatial AVT), AudioCaps [43] (AT), and MS-COCO [44] (VT). Using Binaural only yields lower performance due to weaker semantic alignment. Training all data in a single stage improves scene (34.7→35.2) and grounding (19.6→19.9) but degrades spatial retrieval (38.4→36.5), as non spatial data dilutes spatial supervision. A two-stage schedule in which we train on all data for global alignment and then fine tune on Binaural data for better spatial grounding achieves the best balance, preserving spatial performance while improving overall alignment.

Contribution of Binaural Spatial Cues. We average the left and right channels and duplicate the mono waveform across both inputs, keeping the backbone, architecture, training schedule, and seed fixed. This removes interaural differences while largely preserving semantic content. We evaluate hard A↔V/A↔T retrieval, audio spatial retrieval, and L/R grounding to assess spatial discrimination under this control. Mono input reduces hard A↔T R@1 by 5.3 points, audio spatial R@1 by 5.5, mAP by 5.7, and L/R accuracy by 11.9 (Table 2). The modest change in Hard A↔V retrieval reflects its more holistic scene-matching nature: mono audio preserves the semantic content needed to associate an audio clip with its corresponding visual scene. In contrast, the substantial and consistent drops on audio-conditioned spatial retrieval and grounding, where semantic identity alone is insufficient, directly show that SceneBind extracts and uses binaural spatial cues rather than merely regressing to visual priors. Appendix Sec. E provides per-attribute analysis.

Matching Policy Justification. Global matching captures holistic scene semantics and is strong for semantic retrieval, but lacks spatial precision. Adding slots improves all metrics as shown in Table 4, especially Spatial retrieval (25.4 to 38.4), showing the need for object-centric slots which provide the missing spatial location signal. Slot-only matching preserves Spatial performance but loses scene context, causing large scene retrieval accuracy drops. The slot-only ablations further justify our *Best match* design. In particular, Hungarian one-to-one assignment improves T→A/V when clean GT clauses query predicts slots, but degrades A/V→T by forcing noisy predicted slots to match GT clauses. Removing confidence mainly degrades prediction-to-prediction (A↔V) and prediction-to-GT (A/V→T) matching since noisy low-quality slot pairs are no longer down-weighted.

Object-Centric Slot Scaling. We analyze how the number of object-centric slots affects semantic-spatial performance (Fig. 3). Beyond a moderate number of slots (i.e., > 10), performance saturates, suggesting that a small set of $\sim$10 is sufficient to capture scene-level structure. Visual tasks benefit from more slots (peaking around 25) due to higher object density, while audio which typically has fewer concurrent events peaks with fewer slots (around 5). For zero-shot audio-visual localization (Sec. 4.7), more slots provide better candidate regions and improve coverage. The slot count specifies maximum capacity, while confidence gating suppresses unused slots.

Efficiency. SceneBind's joint audio and visual path takes 13.09 ms on one H200 at batch size one, using a 256×256 frame, two-second binaural audio, and 50 slots (Appendix Sec. C.5).

4.6 Qualitative Results

Figure 4 compares SceneBind with the strongest zero-shot baseline, ImageBind, and the strongest fine-tuned baseline, M2D-SigLIP2*. In A→V retrieval, SceneBind ranks the GT park scene first

Table 3: Object grounding with per-attribute spatial accuracy.

Mod.	Acc ↑	Azi ↑	L/R ↑	Elev ↑	Dist ↑
A	19.4 ± 0.7	39.7 ± 1.4	62.5 ± 0.8	83.5 ± 0.5	58.4 ± 0.2
V	20.0 ± 0.8	36.2 ± 0.8	67.6 ± 0.5	76.2 ± 0.2	65.7 ± 0.4

Table 4: SceneBind Matching ablations. Global and slot matching are complementary; slot-only ablations support our design (*Best match*). For all metrics, the higher, the better.

Strategy	AV	T→A/V	A/V→T	Spatial
Global	21.3	36.1	34.2	25.4
Global+Slot	**21.8**	**42.4**	**39.9**	**38.4**
Slot-only:				
Best match	11.5	18.6	17.3	38.4
Hungarian	10.9	21.1	6.9	38.4
w/o conf	9.5	22.8	15.7	38.4

Table 5: Training objectives ablations. Object semantic loss is key for grounding, intra scene contrastive loss for discrimination, and batch level semantic-spatial contrastive loss for spatial retrieval.

$\mathcal{L}_{\text{sem}}$	$\mathcal{L}_{\text{inst}}$	$\mathcal{L}_{\text{batch}}$	Scene ↑	Spatial ↑	Ground ↑
✓	✓		35.5	36.5	**19.6**
✓		✓	34.3	31.1	16.1
	✓	✓	**36.0**	35.2	8.8
✓	✓	✓	34.7	**38.4**	**19.6**

Table 6: Training schedule and data. Two-stage training (All → Binaural) preserves spatial performance while improving overall alignment.

Stage 1	Stage 2	Scene ↑	Spatial ↑	Ground ↑
Binaural	–	33.8	35.2	19.5
All	Binaural	34.7	**38.4**	19.6
All	All	**35.2**	36.5	**19.9**

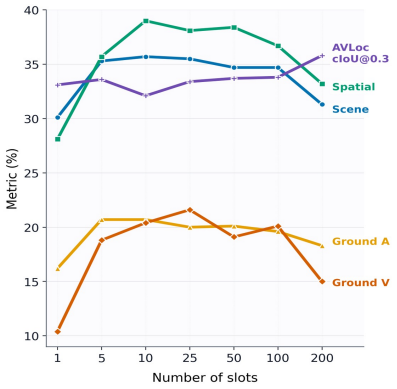

Figure 3: **Object slot scaling.** Performance saturates or drops beyond a moderate number; ∼10 slots suffice.

Table 7: **Zero-shot Sphere360 hard-pool retrieval.** Scene-level retrieval under identical semantics but varying spatial configurations. SceneBind reports three-seed means ± sample std and significantly improves overall and text-related retrieval accuracy. For all metrics, the higher, the better.

Audio–Text	Visual–Text	AV	VT	AT	Avg
AudioCLIP [11]	AudioCLIP [11]	11.2	16.0	4.9	10.7
SpatialCLAP [10]	SigLIP2 [5]	13.6	19.9	1.9	11.8
LAION-CLAP [6]	SigLIP2 [5]	17.5	19.9	8.7	15.4
M2D-CLAP [7]	SigLIP2 [5]	18.4	19.9	<u>10.2</u>	16.2
ImageBind [8]	ImageBind [8]	15.0	18.0	5.3	12.8
M2D-CLAP* [7]	SigLIP2* [5]	<u>19.9</u>	20.4	8.3	16.2
ImageBind* [8]	ImageBind* [8]	**23.3**	<u>22.8</u>	<u>10.2</u>	<u>18.8</u>
SpatialCLAP* [10]	SigLIP2* [5]	19.4	19.9	9.7	16.3
SceneBind	**SceneBind**	**23.3** ± 2.7	**30.9** ± 1.6	**27.7** ± 2.1	**27.3** ± 1.7

and retrieves candidates with similar layouts, while the baselines capture coarse semantics but miss key spatial cues such as "a child crying on the right". SceneBind Matching identifies cues such as "bird sounds" and "a child crying on the right" from object-slot predictions, while the object score (o) complements the global score (g). These balance matches of spatial objects with overall scene context. In T→A and T→V retrieval, SceneBind also matches queried objects with their spatial attributes; in the scooter example, the correct image is selected by a higher slot score despite a lower global score. These examples show SceneBind's ability to align both semantics and locations.

4.7 Zero-shot Downstream Application: Egocentric Spatial Audio-Visual Localization

SceneBind generalizes to egocentric AV localization without task-specific training (Table 8), achieving strong cIoU and AUC, demonstrating effective transfer of semantic-spatial representations.

As shown in Fig. 5, the global semantic query induces an attention map over visual patches, serving as a region-of-interest prior. We perform cross-modal matching between audio and visual object-centric slots using semantic and spatial consistency, and select the best pair. The predicted spatial attributes (azimuth, elevation) define a Gaussian spatial prior, which is fused with the attention map to refine localization, yielding the final sounding region. More details are described in the appendix.

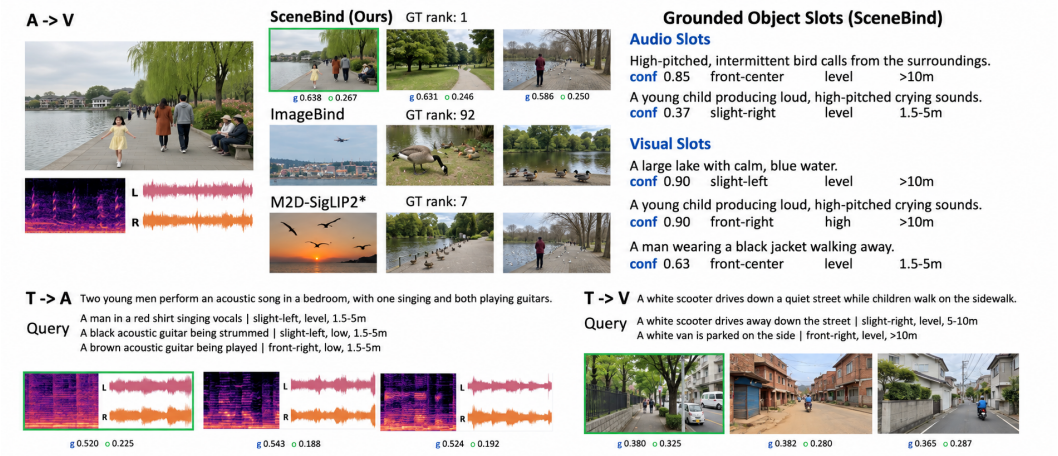

Figure 4: **Qualitative results.** SceneBind retrieves semantically and spatially aligned scenes across audio, vision, and text. We show top-3 candidates with global similarity (g), object-centric matching score (o), and grounded slots with confidence and spatial labels.

Table 8: **Ego AV localization.** SceneBind (AV) uses binaural-only training, while SceneBind uses full multi-source training; both are zero-shot and outperform prior finetuned methods, with full training further improving performance.

Method	cIoU ↑ @0.2	cIoU ↑ @0.3	cIoU ↑ @0.4	AUC ↑
Attention [48]	7.12	–	–	6.42
STM [49]	12.10	7.64	4.01	8.87
Hardway [50]	24.51	13.55	6.10	13.38
SSPL [51]	13.62	8.10	4.45	9.56
Mix [52]	26.01	15.25	9.90	15.39
SSS [53]	9.76	9.89	10.10	7.50
AVLoc [47]	38.71	19.42	10.51	18.38
SceneBind (AV)	<u>43.94</u>	<u>26.34</u>	<u>13.91</u>	<u>20.80</u>
SceneBind	**52.65**	**33.73**	**19.47**	**24.39**

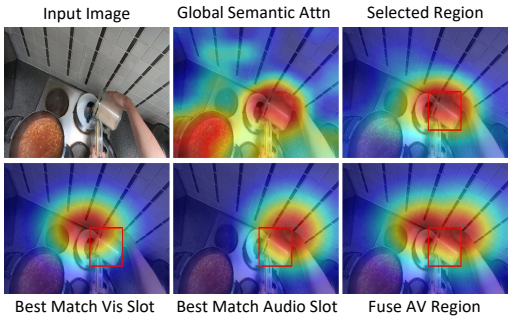

Figure 5: **AV localization visualization.** Global semantic attention proposes sounding regions, refined by spatial consistency of SceneBind slots.

5 Conclusion

We presented **SceneBind**, a unified omni-modal representation for joint semantic and spatial representation across vision, audio and language. By combining global semantics with object semantic-spatial slots, SceneBind models *what* and *where*, enabling spatial-semantic alignment for cross modal retrieval and grounding. We introduce a curated spatially aligned binaural dataset, a joint semantic-spatial training protocol, and a unified matching policy integrating global and object representations. SceneBind achieves strong performance along with effective zero-shot generalization.

Limitations and Future Works. SceneBind's spatial classification heads predict coarse distance, azimuth, and elevation bins rather than exact 3D positions. SceneBind may benefit from scaling real world spatially aligned multimodal data with richer in-the-wild annotations, and from extending to longer temporal windows to model object motion, long-range scene dynamics, and temporally consistent semantic-spatial reasoning. Additional discussion is provided in the Appendix.

6 Acknowledgment

The authors (MC,ES) acknowledge the partial support of HDR Institute: Accelerated AI Algorithms for Data-Driven Discovery (A3D3) National Science Foundation grant PHY-2117997, National Science Foundation grant EFRI-BRAID-2223495 and National Science Foundation grant PFI-TT-2414896. The authors also acknowledge the partial support by the Departments of Electrical Computer Engineering and Applied Mathematics. The authors are thankful to the eScience Center at the University of Washington.

References

[1] Jihan Yang, Shusheng Yang, Anjali Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in Space: How Multimodal Large Language Models See, Remember and Recall Spaces. *arXiv preprint arXiv:2412.14171*, 2024.

[2] Mingfei Chen, Zijun Cui, Xiulong Liu, Jinlin Xiang, Caleb Zheng, Jingyuan Li, and Eli Shlizerman. SAVVY: Spatial awareness via audio-visual LLMs through seeing and hearing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[4] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, 2023.

[5] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim M. Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier H'enaff, Jeremiah Harmsen, Andreas Steiner, and Xiao-Qi Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *ArXiv*, abs/2502.14786, 2025.

[6] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[7] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, Masahiro Yasuda, Shunsuke Tsubaki, and Keisuke Imoto. M2d-clap: Masked modeling duo meets clap for learning general-purpose audio-language representation, 2024.

[8] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[9] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022.

[10] Kentaro Seki, Yuki Okamoto, Kouei Yamaoka, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Spatial-clap: Learning spatially-aware audio–text embeddings for multi-source conditions. *arXiv preprint arXiv:2509.14785*, 2025.

[11] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021.

[12] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

[13] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[14] Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning, 2023.

[15] Sourav Garg, Krishan Rana, Mehdi Hosseinzadeh, Lachlan Mares, Niko Sünderhauf, Feras Dayoub, and Ian Reid. Robohop: Segment-based topological map representation for open-world visual navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4090–4097. IEEE, 2024.

[16] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. *Robotics: Science and Systems (RSS)*, 2023.

[17] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024.

[18] Yuncong Yang, Han Yang, Jiachen Zhou, Peihao Chen, Hongxin Zhang, Yilun Du, and Chuang Gan. 3d-mem: 3d scene memory for embodied exploration and reasoning, 2024.

[19] Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, and Shuqiang Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 15625–15636, 2023.

[20] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[21] Hanlin Chen, Fangyin Wei, and Gim Hee Lee. Chatsplat: 3d conversational gaussian splatting. *arXiv preprint arXiv:2412.00734*, 2024.

[22] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024.

[23] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024.

[24] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *ArXiv*, abs/2303.05499, 2023.

[25] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding dino 1.5: Advance the "edge" of open-set object detection. *ArXiv*, abs/2405.10300, 2024.

[26] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, Lecture Notes in Computer Science, pages 213–229. Springer, 2020.

[27] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[28] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1184–1193, 2021.

[29] Moitreya Chatterjee, Narendra Ahuja, and Anoop Cherian. Learning audio-visual dynamics using scene graphs for audio source separation. In *Advances in Neural Information Processing Systems*, 2022.

[30] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10565–10574, 2023.

[31] Yuxin Ye, Wenming Yang, and Yapeng Tian. Lavss: Location-guided audio-visual spatial audio separation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5508–5519, 2024.

[32] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[33] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33:10077–10087, 2020.

[34] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2744–2753, 2021.

[35] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):34–48, 2018.

[36] David Diaz-Guerra, Antonio Miguel, and Jose R Beltran. Robust sound source tracking using srp-phat and 3d convolutional neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:300–311, 2020.

[37] Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. Bat: Learning to reason about spatial sounds with large language models. *International conference on machine learning*, 2024.

[38] Bhavika Suresh Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, YELENA MENYAYLENKO, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. Learning spatially-aware language and audio embeddings. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[39] Artem Dementyev, Wazeer Zulfikar, Sinan Hersek, Pascal Getreuer, Anurag Kumar, and Vivek Kumar. Phasecoder: Microphone geometry-agnostic spatial audio understanding for multimodal llms, 2026.

[40] Hyeonggon Ryu, Joon Son Chung, and David Harwath. Hear you are: Teaching llms spatial reasoning with vision and spatial sound. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 38606–38615, 2026.

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[42] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955.

[43] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[45] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[46] Huadai Liu, Tianyi Luo, Qikai Jiang, Kaicheng Luo, Peiwen Sun, Jialei Wan, Rongjie Huang, Qian Chen, Wen Wang, Xiangtai Li, Shiliang Zhang, Zhijie Yan, Zhou Zhao, and Wei Xue. Omniaudio: Generating spatial audio from 360-degree video. *ArXiv*, abs/2504.14906, 2025.

[47] Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Egocentric audio-visual object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22910–22921, 2023.

[48] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[49] Sizhe Li, Yapeng Tian, and Chenliang Xu. Space-time memory network for sounding object localization in videos. *arXiv preprint arXiv:2111.05526*, 2021.

[50] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021.

[51] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3222–3231, 2022.

[52] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022.

[53] Hyeonggon Ryu, Seongyu Kim, Joon Son Chung, and Arda Senocak. Seeing speech and sound: Distinguishing and locating audio sources in visual scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13540–13549, 2025.

[54] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[55] Bill Gardner and Keith Martin. Hrtf measurements of a kemar dummy-head microphone. 1994.

A Appendix Overview

In this appendix, we provide supplementary details and analyses that complement the main paper.

- **Dataset Curation** (Sec. B): We describe the data sources, annotation pipeline, verification process, balancing strategy, benchmark statistics, and annotation agreement.
- **Implementation Details** (Sec. C): We provide architectural, training, and inference-time matching details for SceneBind, including inference efficiency measurements and spatial supervision ablations.
- **Baseline Fine-tuning** (Sec. D): We describe how retrieval baselines are trained and evaluated under the same data and benchmark settings.
- **Balanced Spatial Evaluation and Binaural Cue Ablation** (Sec. E): We describe the spatial evaluation metrics and analyze binaural cues through a controlled mono comparison and per-bin elevation results.
- **Zero-shot Audio-Visual Localization Details** (Sec. F): We detail the zero-shot audio-visual localization procedure and evaluation protocol.
- **Discussion** (Sec. G): We discuss SceneBind's semantic-spatial representation, learning from weak spatial supervision, and complementary global and object-level structure.
- **Limitations and Future Works** (Sec. H): We discuss current limitations and future research directions.
- **Broader Impacts** (Sec. I): We discuss potential societal impacts and safeguards for SceneBind.

We also provide an interactive qualitative viewer in the supplementary folder (`index.html`) for cross modal retrieval and object grounding. It covers all six scene retrieval directions and displays query and retrieved samples with images, binaural audio, and text annotations. We recommend spatial audio playback for audio examples.

B Dataset Curation

SceneBind is trained and evaluated on in-the-wild audio-visual data with aligned semantic and spatial annotations. To support semantic-spatial learning, we curate data from multiple real-world sources and convert them into a unified format containing paired audio-visual samples, object-level text clauses, and spatial labels. This section provides details of the dataset curation process including the data sources, processing pipeline and annotation procedure used for training and evaluation.

B.1 Data Source

Binaural Dataset. We construct the *Binaural* training corpus from in-the-wild binaural-audio videos sourced from publicly available platforms. Each video is cropped to samples where each sample contains a 2-second binaural clip, a representative image frame, and event annotations with semantic and spatial descriptions. The corpus is designed to support omni-modal spatial learning from in-the-wild scenes.

Sphere360 Dataset. We additionally construct the *Sphere360* subset from YouTube 360° videos with first-order ambisonic (FOA) audio [46]. Its 360° coverage and FOA audio enable rendering each scene from arbitrary listener orientations, producing the spatially augmented candidate pool for each query sample. In our experimental setting, Sphere360 serves exclusively as a zero-shot evaluation benchmark, where candidates share the same semantics but differ in spatial configuration, testing the model's ability to discriminate fine-grained spatial structure under semantic ambiguity.

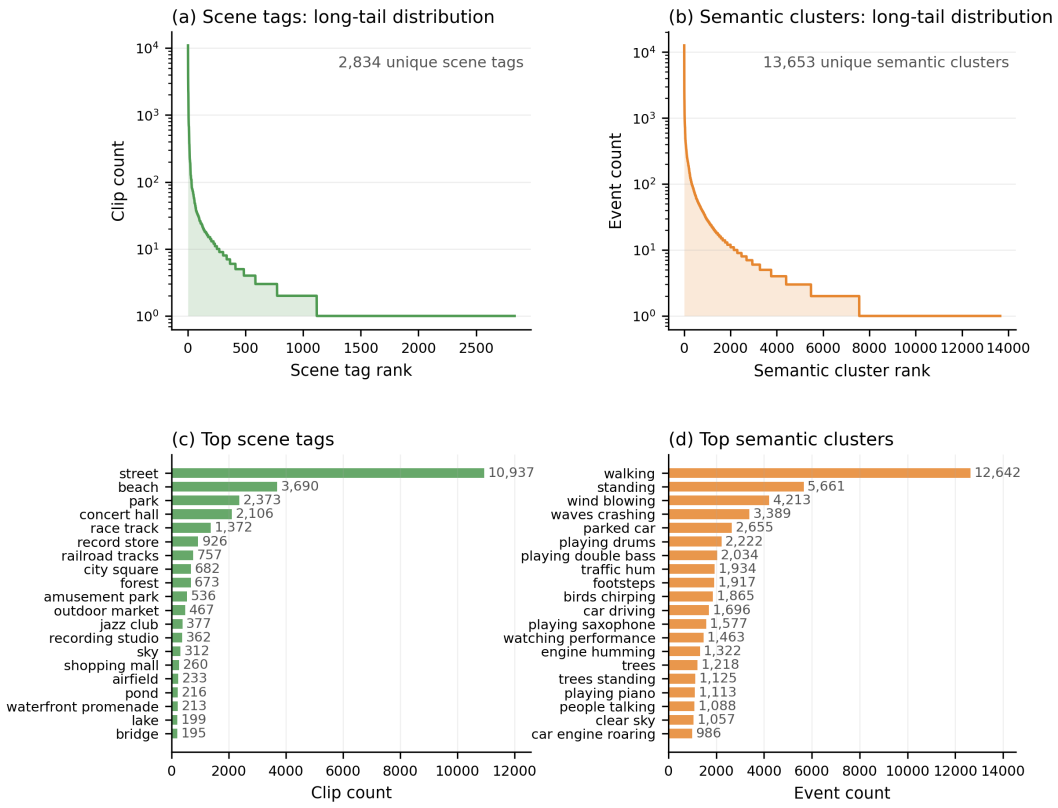

Figure 6: Long-tail distributions of scene tags and semantic event clusters in the Binaural dataset. The top row shows rank-frequency curves for scene tags and semantic clusters on a log-scaled count axis. The bottom row reports the 20 most frequent scene tags and semantic clusters. The scene tags span urban, transportation, performance, and outdoor environments, while the semantic clusters cover music, human motion, vehicles, and ambient events. This distribution shows that the curated data contains diverse real-world scenes and a broad range of sound-producing objects and actions.

B.2 Pipeline

We construct the binaural training data with a multi-stage pipeline designed to identify audio-visual events with plausible spatial annotations from source binaural video corpus. The pipeline starts from source in-the-wild videos, filters for usable binaural audio and non-360 content, selects diverse 10-second candidate windows, annotates 2-second event segments, verifies audio-visual grounding with independent encoders, and finally applies data balancing and human review. Figure 7 summarizes the resulting data funnel.

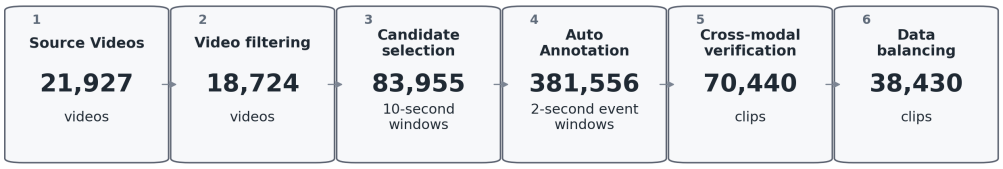

Figure 7: Binaural data construction funnel. The pipeline filters source videos, selects diverse candidate windows, annotates short event segments, verifies audio-visual grounding, and applies data balancing to produce the released training set.

Specifically, starting from 21,927 source videos, the pipeline retains 18,724 after video-level filtering, selects 83,955 candidate 10-second windows from 8,015 videos via per-clip scoring and Bayesian Gaussian Mixture Models (BayesGMM) diversity selection, annotates 381,556 candidate 2-second event windows, verifies 70,440 clips with at least one audio-visual event, and produces 56,947 clips after spatial balancing.

Candidate Construction. For the binaural corpus, we first fetch video metadata and retain videos that satisfy deterministic quality checks: landscape orientation, resolution at least 1920 pixels in width or 1080 pixels in height, source information explicitly indicating binaural-microphone capture, and non-360 projection. Each retained video is then divided into short candidate windows in 2 seconds and scored using three complementary signals from pretrained audio-visual models and binaural audio statistics. First, VGGish [54] features measure acoustic informativeness and are used to filter out uninformative audio clusters. Second, ImageBind [8] measures audio-visual alignment, retaining clips whose audio and visual embeddings are sufficiently similar. Third, binaural-cue strength is estimated from frequency-weighted inter-channel level differences and inter-channel coherence. We keep candidates that pass the acoustic cluster filter, have strong audio-visual alignment, and contain reliable binaural cues. Accepted short clips are then grouped into 10-second annotation windows, preserving temporal context for subsequent event annotation and verification.

Candidate construction for Sphere360 is deterministic. We identify each 360° clip as either Equirectangular or Equi-Angular Cubemap (EAC) from its metadata, and render four perspective views at 90° azimuth intervals with zero pitch and a 90° field of view. For each view, we generate the corresponding binaural audio by rotating the first-order ambisonic signal to the view direction and decoding it with the MIT KEMAR HRTF [55] to obtain stereo audio. This produces aligned audio-visual samples with known relative viewing directions. Sphere360 contains real ambisonic recordings with natural sources, acoustics, and noise. Rotating and decoding these recordings changes listener orientation without synthesizing dry sources, room acoustics, or scene content.

Annotation. For the binaural video candidates, we use Gemini [45] to generate initial event proposals from each 10-second video window sampled at 2 fps. Although Gemini does not directly perceive binaural spatial cues, it provides strong visual recognition and temporal audio-visual event understanding. ImageBind and CLAP filter unsupported event-source associations. We guide it to identify audible events, associate them with visible entities when possible, and estimate the corresponding camera-coordinate spatial labels from the visual evidence. The prompt asks for atomic 1–2 second events, assigns each event to one of three modality types (`audio_visual`, `audio_only`, or `visual_only`), and returns schema-constrained JSON with azimuth, elevation, and distance annotations. This produces plausible spatial annotations for each short event window, which are later filtered by cross-modal verification and spatial balancing. Specifically, we define spatial labels in camera coordinates. Azimuth uses seven coarse directional regions: hard left $[-90^\circ, -60^\circ)$, left $[-60^\circ, -30^\circ)$, slight left $[-30^\circ, -15^\circ)$, front $[-15^\circ, 15^\circ]$, slight right $(15^\circ, 30^\circ]$, right $(30^\circ, 60^\circ]$, and hard right $(60^\circ, 90^\circ]$. Elevation uses five vertical regions: high above $[30^\circ, 90^\circ]$, above $[10^\circ, 30^\circ)$, level $[-10^\circ, 10^\circ]$, below $[-30^\circ, -10^\circ)$, and down below $[-90^\circ, -30^\circ)$. Distance is annotated with four ranges: touching/very close $[0, 1.5)$ m, close $[1.5, 5)$ m, medium $[5, 10)$ m, and far $[10, \infty)$ m. We run Gemini deterministically with temperature 0. The full prompt for the binaural subset is shown in Fig. 8.

For Sphere360, we use a three-round protocol to ensure that annotation accounts for multiple possible viewing directions. 1) In Round 1, Gemini annotates each of the four cardinal views independently with a consistent prompt as shown in Fig. 8, and identifies the primary view with the strongest audio-visual correspondence. 2) In Round 2, each candidate event is revisited using five auxiliary neighbor views centered around the primary view, with yaw offsets $\{0^\circ, \pm 30^\circ, \pm 60^\circ\}$. We input the Round-1 annotations with the neighbor views to Gemini to confirm or revise each event across these viewpoints. 3) In Round 3, we merge cross-view annotations for the same event, retaining the entry whose view places the event closest to the frontal direction, i.e., with the smallest absolute azimuth. The merged event can then be propagated to the rendered views by updating its relative azimuth according to each view direction. This produces augmented audio-visual samples with consistent semantic labels and view-dependent spatial annotations across the 360-degree views.

```
Analyze the provided 10-second clip using both AUDIO and VISUAL cues.
Your goal is to identify audio-visual events where visible objects produce clearly
synchronized sound, and group them by the physical object responsible.
Then, to provide complete scene context, enumeratively identify visual-only and
audio-only objects present at the same timestamps of audio_visual events.

1. Event Modality Constraints
- audio_only: if you can hear the object, but it is not in the field of view.
- visual_only: if visible, but it is silent or not synchronized with any sound.
- audio_visual: if visible AND the sound is clearly synchronized with its motion AND
there is strong audio-visual correspondence. High precision required.

2. Annotation Rules (Strict)
- Timestamps: Must be relative to clip start (0s to 10s).
- Duration Constraint: Events must be short atomic instances. Duration must be EXACTLY
1 or 2 seconds.
- Correct Format: [1,3), or [2,3) relative to clip start; Incorrect Format:
[00:01:21-00:01:23) absolute video timestamps.
- Split Rule: If an event lasts longer than 2 seconds, split into multiple events; for
each split, consider the event annotation, angle and distance without any bias from
previous or subsequent splits.
- For audio_visual event: When in doubt, leave it out. False positives are much worse
than false negatives.
- Valid Duration: When there is at least one valid audio_visual event during that 1 or
2 seconds; only annotate an audio_only or visual_only event if the duration is valid.
- For visual_only and audio_only event: Be as enumerative as you can in the valid
duration.

3. Spatial Definitions (Camera Coordinates)
Azimuth (Horizontal): Center is 0 deg; Negative is left; Positive is right.
- Hard Left: [-90, -60)
- Left: [-60, -30)
- Slight Left: [-30, -15)
- Front: [-15, 15]
- Slight Right: (15, 30]
- Right: (30, 60]
- Hard Right: (60, 90]

Elevation (Vertical): Level is 0 deg; Negative is Down; Positive is Up.
- High Above: [30, 90]
- Above: [10, 30)
- Level: [-10, 10]
- Below: [-30, -10)
- Down Below: [-90, -30)

4. Field Descriptions
- semantic_tag: A concise semantic tag for the event, e.g., 'car honking', 'playing
piano'.
- semantic_anno (6 to 8 words): Describe WHAT the object is doing/being. Strictly
Forbidden: Directional words.
- spatial_anno (about 8 words): Describe WHERE the object is. Strictly Forbidden:
Naming the object, complex action.
- combined_anno (about 16 words): Synthesize both semantic and spatial details
naturally.
- reason: Provide the reasoning behind your estimations for timestamp, semantic_tag,
direction, height, and distance based on audio and visual cues.
```

Figure 8: Audio–Visual event and spatial annotation prompt for binaural videos via Gemini.

Verification and Balancing. Gemini provides useful event proposals, but its audio-visual associations and spatial estimates can be noisy. To mitigate hallucination, we add a further verification stage to retain short windows whose event-source associations are supported by both the visual frame and the audio segment. This step filters hallucinated events, downgrades partially supported events to the appropriate modality type, and balances the spatial distribution of the final data.

Table 9: Event disposition after independent ImageBind (IB) and CLAP verification.

Modality	Threshold pass	Result
audio-visual	ImageBind & CLAP	keep AV
audio-visual	ImageBind	downgrade to VO
audio-visual	CLAP	downgrade to AO
audio-visual	neither	drop
audio-only	CLAP	keep AO
audio-only	fail	drop
visual-only	ImageBind	keep VO
visual-only	fail	drop

Each annotated event is verified with pretrained foundation models that are independent of Gemini. Specifically, we compute an ImageBind frame-text score $s^{\mathrm{IB}}_{v\to t}$ and a CLAP [6, 7] audio-text score $s^{\mathrm{CLAP}}_{a\to t}$. An audio-visual event is retained as audio-visual only when both checks pass ($s^{\mathrm{IB}}_{v\to t} \geq 0.10$ and $s^{\mathrm{CLAP}}_{a\to t} \geq 0.05$). If only the visual check passes, the event is downgraded to visual-only; if only the audio check passes, it is downgraded to audio-only; if neither passes, it is discarded. We summarize the verification rules with respect to different original annotated modality in Tab. 9.

After verification, we select the best 2-second event window with the strongest verified event evidence from each 10-second candidate, then choose the representative frame within that window by maximizing ImageBind alignment with the verified audio-visual event labels. We define the selected 2-second audio and its paired image with event and spatial annotations as a clip. It produces 70,440 verified clips with at least one supported audio-visual event. For Binaural dataset, to reduce front-facing bias, we down-sample clips whose primary audio-visual event is centered at azimuth $= 0°$ with a 50% keep rate. This step produces a spatially balanced corpus of 56,947 clips, yields a more uniform directional distribution as shown in Figure 9, and discourages models from relying on a front-center shortcut. Sphere360 does not use the spatial balancing step because its $360°$ multi-view augmentation already covers all directions.

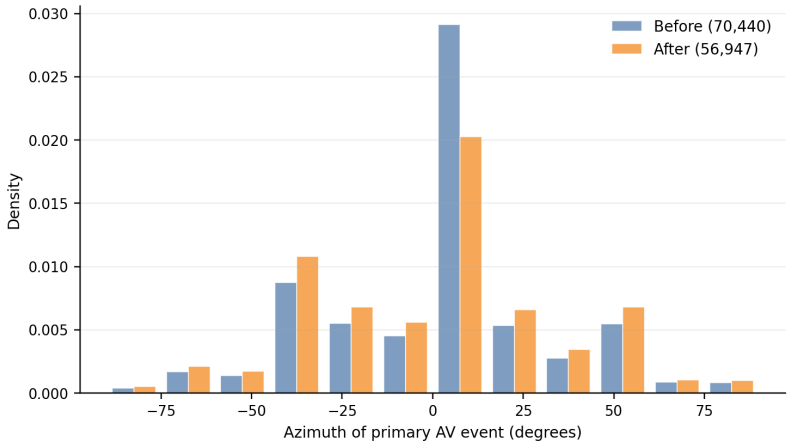

Figure 9: Azimuth distribution of the primary audio–visual event per clip, before (70,440 clips) and after (56,947 clips) spatial balancing. Spatial balancing down-samples the azimuth $= 0°$ bin at a 50% keep rate, producing a more uniform directional distribution.

We then create train, validation, and test splits at the video level with an 80%/5%/15% ratio, stratifying by channel and semantic cluster to reduce train-test leakage. Specifically for training set, we further balance the training clips to reduce redundancy across scene environment tags, event tags, and

YouTube channels, improving diversity. After this diversity balancing, the final training set contains 38,430 binaural clips.

B.3 Human Verification and Interface

We add a human review pass for the binaural benchmark and the zero-shot Sphere360 benchmark. Four reviewers inspect candidate test clips and remove examples with ambiguous content, weak audio-visual correspondence, inaccurate spatial labels, or near-duplicate scenes. When possible, reviewers correct minor spatial-label errors. This final step improves benchmark reliability by retaining clips with meaningful events and reliable semantic-spatial annotations.

For the 1,066-clip Binaural benchmark, the final annotations contain 1,386 audio-relevant and 2,682 visual-relevant objects. Reviewers correct spatial bins using visible camera-relative position, motion consistency, scene geometry, familiar object scale, relative depth ordering, and binaural directional cues. The resulting labels are human-verified categorical judgments.

Review Interface. We build a review interface for clip-level inspection (Fig. 10). The interface presents the representative image, binaural audio, scene description, object clauses, and associated spatial labels. Reviewers use it to verify audio-visual correspondence, label sample quality, discard invalid examples, and correct spatial annotations when needed. Each reviewer works independently, and the reviewed labels are synchronized in the backend and merged into the final benchmark annotations.

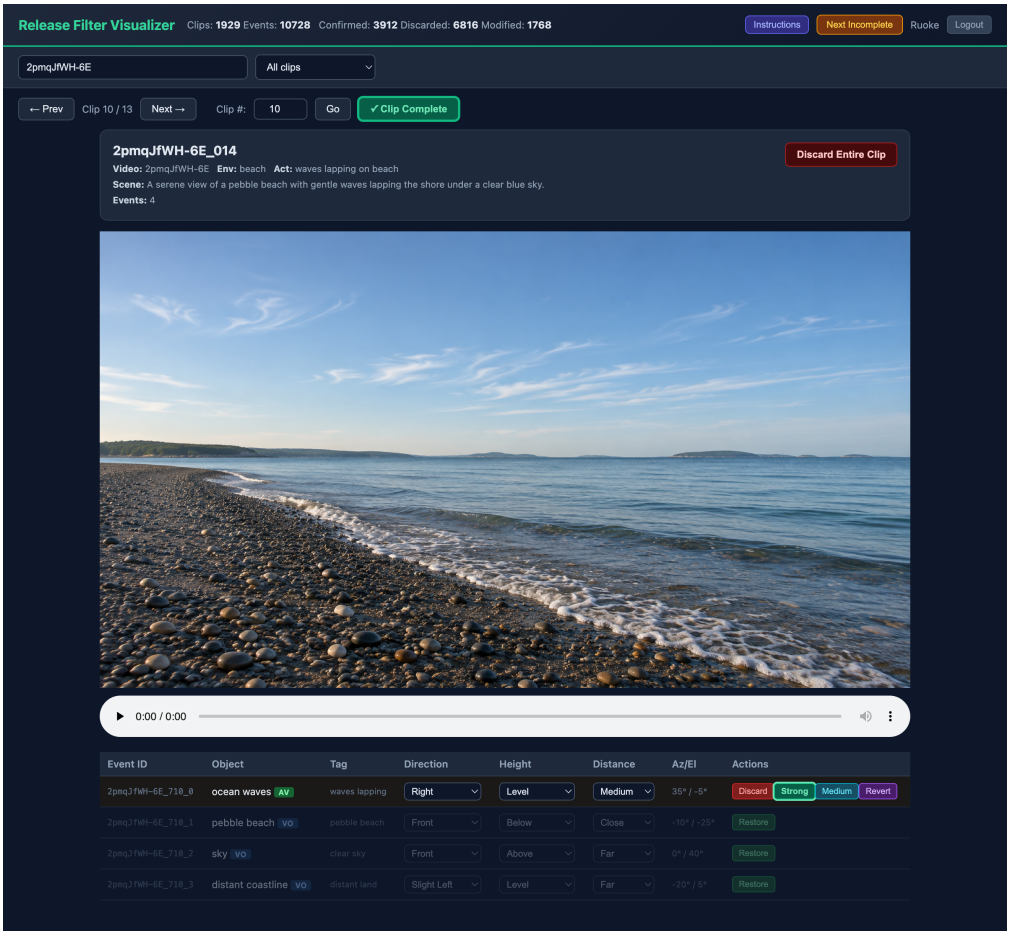

Figure 10: Human verification interface for reviewing candidate benchmark clips. It allows reviewers to inspect the representative image, binaural audio, scene description, object clauses, and spatial labels, then discard invalid samples or correct annotations when needed.

Review Guidelines. Reviewers first decide whether a candidate clip should be kept. A clip is discarded if it fails any of three criteria: 1) weak relevance, where the audio, visual evidence, and semantic description do not describe the same event; 2) object or event ambiguity, where the annotation cannot be resolved to a clear, unique, and non-trivial entity in the clip; or 3) low sample quality, such as unclear imagery, overly noisy audio, or the absence of meaningful events.

For clips that pass this clip-level check, reviewers then inspect each annotated object. If the semantic annotation is correct and the object has meaningful spatial attributes, reviewers verify and, when necessary, correct its azimuth, elevation, and distance labels. If the object is not semantically valid, spatially meaningful, or uniquely identifiable, it should be removed from the clip.

We ask reviewers to assign a quality label for each object of the retained objects. *Strong* indicates that the audio, visual evidence, semantic and spatial annotation clearly agree and describe a meaningful object or event. *Medium* indicates that the annotation is mostly consistent but contains minor ambiguity or uncertainty.

Finally, reviewers remove near-duplicate clips. When multiple clips show nearly identical scene instances in image, audio, or text, we retain only one clip, prioritizing the sample with clearer events, higher object quality, and greater semantic-spatial diversity.

Annotation Agreement. We compare initial automatic labels with final human-verified labels and measure pairwise agreement among the four reviewers before final verification (Table 10). Most disagreements are local, consistent with spatial ambiguity and bin-boundary effects. This motivates Gaussian-smoothed supervision over neighboring spatial bins.

Table 10: **Spatial annotation agreement (%).** Automatic-human agreement compares initial automatic labels with final human-verified labels. Human-human agreement compares reviewer annotations before final verification.

Comparison	Agreement	Azimuth	Elevation	Distance
Automatic–human	Exact	72.40	90.10	83.80
Automatic–human	Within one bin	88.98	99.26	98.72
Human–human	Pairwise exact	79.64	88.73	85.12
Human–human	Pairwise within one bin	97.75	98.03	99.83

B.4 Final Benchmark Details

After human verification, we compile two evaluation benchmarks from the verified clips: the binaural benchmark and the Sphere360 benchmark. This section summarizes the final split statistics and query-pool construction used for evaluation.

Binaural Scene Retrieval. The binaural benchmark contains 1,066 clips in total. For scene-level retrieval, the pool size can differ by modality because some clips do not contain valid objects for every modality. The audio-visual pool contains 1,040 paired samples, while the audio-text and visual-text retrieval pools contain 942 and 1,046 clips, respectively.

For general audio-visual retrieval, all paired audio-visual samples are used as queries. For text-based scene retrieval, we focus on non-trivial foreground queries: audio-text queries require more than one audio-relevant object, i.e., audio-visual plus audio-only objects, and visual-text queries require more than one visual-relevant object, i.e., audio-visual plus visual-only objects. This filtering avoids degenerate single-object scenes where the text query does not test object-level disambiguation. Consequently, the resulting benchmark contains 1040 audio-visual scene queries, 325 audio-text scene queries, and 757 visual-text scene queries.

Binaural Spatial Retrieval. Binaural spatial retrieval uses object-level text queries. Starting from the audio-relevant and visual-relevant object annotations, we retain a query only when its semantic cluster forms a valid hard pool: the pool must contain at least one sample with matching quantized azimuth, elevation, and distance labels, and its size must fall within the predefined range (e.g., $[5, 20]$). This yields 745 audio-relevant queries and 782 visual-relevant queries, for 1,527 total queries.

Spatial mAP ranks candidates from the full retrieval pool as the scene retrieval task using the joint semantic-spatial object score, and treats samples with matching quantized spatial labels as positives.

Spatial-hard R@1 evaluates the same queries in a harder setting: each query is ranked only within its same-cluster hard pool, which contains at most one sample per unique spatial bin. The retained hard pools have mean size 11.75 and median size 10 overall.

Binaural Object Grounding. Object grounding evaluates the model's predicted object-centric slots against all modality-relevant annotations in the 1,066 binaural clips. The benchmark includes 1,386 audio-relevant targets from audio-visual and audio-only objects, and 2,682 visual-relevant targets from audio-visual and visual-only objects.

Sphere360 Retrieval. The Sphere360 hard benchmark evaluates spatial disambiguation when objects share similar semantics but appear from different viewing directions. The split contains 97 clips and 4,277 annotated objects, including 553 audio-visual objects. Query objects are restricted to frontal examples with azimuth in $[-90^\circ, 90^\circ]$, while candidate pools are drawn from augmented 360-degree views. The per-query candidate pool has mean size 45.4 and median size 47.

C SceneBind Implementation Details

C.1 SceneBind Encoding Model

Visual and Text Encoder. We use a frozen SigLIP2 [5] visual backbone as the visual feature extractor. For each frame, the backbone produces 256 patch tokens in a 1152-dimensional embedding space, which serve as the visual context tokens. During training, we apply horizontal flip augmentation by using the corresponding flipped visual features and mirroring the azimuth label.

For text encoding, we use the frozen text tower of the same SigLIP2 model. We encode two types of text supervision. First, the scene description is encoded as a single global 1152-dimensional text embedding, which supervises global semantics alignment. Second, each object clause is encoded independently as a 1152-dimensional semantic embedding, which matches with audio and visual object slots. Each object clause is also associated with spatial labels $\mathbf{r}_k = (\theta_k, \phi_k, d_k)$, represented as one-hot targets over the discretized azimuth, elevation, and distance bins.

Audio Semantics Encoder. The binaural waveform is loaded as a two-channel signal and padded or cropped to the fixed audio window. For semantic audio encoding, we average the two channels to obtain a mono waveform and resample it to 16 kHz. A frozen M2D-CLAP audio encoder [7] extracts a sequence of 768-dimensional audio patch tokens. We then use a residual MLP alignment module to project these tokens into the 1152-dimensional SigLIP2 embedding space. The alignment module contains five residual blocks with hidden width 1024 and LayerNorm/GELU nonlinearities. It is pretrained to align M2D-CLAP audio features with SigLIP2 text features on the Binaural training captions, and both the M2D backbone and the alignment module are frozen during SceneBind training. After alignment, the semantic audio representation has shape 310×1152 for each 10-second audio window. These 310 tokens correspond to a 5×62 spectro-temporal patch grid over the M2D log-mel representation, with 5 frequency patches and 62 time patches.

Audio Spatial Encoder. For spatial audio encoding, we use the original binaural waveform. The binaural signal is resampled to 16 kHz and converted into a four-channel time-frequency representation consisting of left/right magnitude and phase features from an STFT with FFT size 1024 and hop size 512. A convolutional frontend processes this representation with three convolutional blocks, Each block uses a 3-by-3 convolution with 64 output channels, ReLU, batch normalization, and max pooling on the frequency axis. SceneBind uses the resulting convolutional feature map as the spatial patch representation. Since the frontend operates on an STFT frequency grid that is not aligned with the M2D patch grid, we average over the remaining frequency dimension to obtain a temporal sequence of 64-dimensional binaural features with frequency-aggregated spatial cues. The learned 2-D convolutions extract local interaural and spectral patterns before frequency pooling. Averaging removes explicit frequency-bin identity while retaining aggregated responses to these patterns. Each feature is projected to the shared 1152-dimensional space and linearly interpolated to length 310, forming spatial audio tokens that match the flattened M2D semantic token sequence.

Semantic Decoder. For each modality, SceneBind uses a query-based semantic decoder. Query embeddings are fixed at inference, while decoded slots remain input-dependent and confidence gating selects active slots. We initialize one learnable global query and $K = 50$ learnable object queries. A one-layer Transformer decoder cross-attends these learnable queries to the modality-specific semantic context tokens. The global-query output is used as the global semantic embedding, while

the object-query outputs define semantic embeddings for predicted object slots. Each object slot is additionally passed through a confidence head, producing a score in $[0, 1]$ that indicates whether the slot corresponds to an active object.

Let H^m denote the semantic context tokens for modality m, and let $Q = [q_{\text{global}}, q_1, \ldots, q_K]$ denote the learnable decoder queries. The semantic decoder computes

$$[\mathbf{s}^m_{\text{global}}, \mathbf{s}^m_1, \ldots, \mathbf{s}^m_K] = \text{Dec}_{\text{sem}}(Q; H^m),$$

where $\mathbf{s}^m_{\text{global}}$ is the global embedding and $\mathbf{s}^m_i$ is the semantic embedding of object slot i.

Spatial Decoder. The spatial decoder predicts spatial attributes for each object slot by conditioning its semantic embedding on modality-specific spatial evidence. Its output is passed through shared spatial projection heads and classifiers for azimuth, elevation, and distance. The classifiers produce distributions $\boldsymbol{\pi}^{m,\theta}_i$, $\boldsymbol{\pi}^{m,\phi}_i$, and $\boldsymbol{\pi}^{m,d}_i$ over 7, 5, and 4 bins (defined in Sec. B.2), respectively; their argmax bins form $\hat{\mathbf{r}}^m_i = (\hat{\theta}^m_i, \hat{\phi}^m_i, \hat{d}^m_i)$ (defined in Sec. B.2).

For the visual branch, semantic object slots serve as decoder queries, while visual context tokens from the visual encoder serve as memory:

$$h^v_i = \text{Dec}_{\text{spa}}(\mathbf{s}^v_i; H^v), \qquad \hat{\mathbf{r}}^v_i = g_{\text{spa}}(h^v_i),$$

where h^v_i is the decoded spatial feature and g_{spa} denotes the shared spatial heads.

For the audio branch, semantic object slots again serve as decoder queries. The decoder memory is the fused audio context, obtained by adding semantic audio tokens and length-aligned spatial audio tokens:

$$H^a_{\text{fused}} = H^a_{\text{sem}} + \widetilde{H}^a_{\text{spatial}}, \qquad h^a_i = \text{Dec}_{\text{spa}}(\mathbf{s}^a_i; H^a_{\text{fused}}), \qquad \hat{\mathbf{r}}^a_i = g_{\text{spa}}(h^a_i).$$

Thus, the semantic slots first form object-level hypotheses, and the spatial decoder then retrieves visual or spatial audio evidence for each hypothesized object. This enables SceneBind to handle variable numbers of objects or events within a scene.

C.2 Training Objective Details

We train SceneBind with a combination of global-level semantic alignment, object-centric semantic-spatial grounding, and cross-modal semantic-spatial contrastive losses.

Global Alignment. Scene-level alignment is trained with a symmetric InfoNCE [41] loss between paired modalities. Let $S^m_{\text{global}} \in \mathbb{R}^{B\times d}$ denote the batch-stacked normalized global embeddings for modality m. For paired modalities m and n, we compute

$$\mathcal{L}^{m,n}_{\text{global}} = \frac{1}{2}\left[\text{CE}\left(\frac{S^m_{\text{global}}(S^n_{\text{global}})^\top}{\tau}, I\right) + \text{CE}\left(\frac{S^n_{\text{global}}(S^m_{\text{global}})^\top}{\tau}, I\right)\right], \qquad \tau = 0.07.$$

This loss is applied whenever paired sample-level supervision is available. For text, global alignment uses the scene-description embedding.

Object Grounding Supervision. After assigning ground-truth object clauses to predicted object slots by bipartite matching, we supervise each matched slot with the semantic, confidence, and spatial losses defined in Sec. 3.3. The semantic loss is a cosine-distance loss between the predicted slot embedding and the matched object-clause embedding. The confidence loss is binary cross-entropy, where matched slots are treated as positives and unmatched slots as negatives. Spatial attributes are supervised with Gaussian-smoothed cross-entropy over discretized azimuth, elevation, and distance bins. We use circular smoothing with $\sigma = 1.0$ for azimuth and non-circular smoothing with $\sigma = 0.5$ for elevation and distance.

Intra-Scene Contrastive Loss $\mathcal{L}_{inst}$. For an audio-visual pair, slots from the two modalities are first matched to the same ground-truth object clauses. Co-matched slot pairs are treated as positives within the scene. If multiple objects are present, we apply a symmetric contrastive loss over the co-matched audio and visual slots, so that each object is closest to its counterpart in the other modality. The same principle is used for audio-text and visual-text grounding: the modality slot matched to an object is treated as a positive pair with the corresponding object-clause text embedding, while other clauses in

the same scene serve as negatives. If only one object is co-matched, the loss reduces to paired cosine alignment.

Cross-Scene Contrastive Loss $\mathcal{L}_{\text{batch}}$. To improve object-level discriminability across scenes, SceneBind applies batch-wise slot contrastive losses over slots matched to ground-truth object clauses. For audio-visual training, audio and visual slots independently matched to the same ground-truth object form positive pairs. We apply symmetric InfoNCE to both semantic slots and spatial-head embeddings:

$$\mathcal{L}^{a,v}_{\text{batch}} = \mathcal{L}_{\text{NCE}}(\mathbf{s}^a, \mathbf{s}^v) + \mathcal{L}_{\text{NCE}}(\mathbf{u}^{a,\theta}, \mathbf{u}^{v,\theta}) + \frac{1}{2}\mathcal{L}_{\text{NCE}}(\mathbf{u}^{a,\phi}, \mathbf{u}^{v,\phi}) + \frac{1}{2}\mathcal{L}_{\text{NCE}}(\mathbf{u}^{a,d}, \mathbf{u}^{v,d}),$$

where $\mathbf{u}^{\theta}$, $\mathbf{u}^{\phi}$, and $\mathbf{u}^{d}$ are normalized spatial-head embeddings before the final classifiers.

For audio-text and visual-text pairs, the cross-scene contrastive loss is semantic-only because text clauses do not provide learned spatial-head features:

$$\mathcal{L}^{a,t}_{\text{batch}} = \mathcal{L}_{\text{NCE}}(\mathbf{s}^a, \mathbf{s}^t), \qquad \mathcal{L}^{v,t}_{\text{batch}} = \mathcal{L}_{\text{NCE}}(\mathbf{s}^v, \mathbf{s}^t).$$

Full Loss. The final objective is a weighted sum of the losses described above. Table 11 summarizes the loss terms and weights used in the main training setting.

Table 11: Loss weights used in the main SceneBind training setting.

Category	Loss term	Weight
Object grounding	Semantic grounding $\mathcal{L}_{\text{sem}}$	1.0
Object grounding	Confidence $\mathcal{L}_{\text{conf}}$	2.0
Object grounding	Azimuth $\mathcal{L}_{\theta}$	1.0
Object grounding	Elevation $\mathcal{L}_{\phi}$	0.5
Object grounding	Distance $\mathcal{L}_{d}$	0.5
Global alignment	Audio-visual $\mathcal{L}^{a,v}_{\text{global}}$	2.0
Global alignment	Audio-text $\mathcal{L}^{a,t}_{\text{global}}$	1.0
Global alignment	Visual-text $\mathcal{L}^{v,t}_{\text{global}}$	1.0
Intra-scene contrastive	Audio-visual $\mathcal{L}^{a,v}_{\text{inst}}$	1.0
Intra-scene contrastive	Audio-text $\mathcal{L}^{a,t}_{\text{inst}}$	1.0
Intra-scene contrastive	Visual-text $\mathcal{L}^{v,t}_{\text{inst}}$	0.5
Cross-scene contrastive	Audio-visual $\mathcal{L}^{a,v}_{\text{batch}}$	2.0
Cross-scene contrastive	Audio-text $\mathcal{L}^{a,t}_{\text{batch}}$	0.5
Cross-scene contrastive	Visual-text $\mathcal{L}^{v,t}_{\text{batch}}$	0.5

C.3 Spatial Supervision Ablation

We evaluate Gaussian-smoothed spatial targets and cross-batch spatial contrastive supervision under the observed annotation regime. All variants are trained on the same automatically labeled training set and evaluated on the same human-verified benchmark (Table 12).

Table 12: **Spatial supervision ablations.** Single-seed spatial retrieval results using automatic training annotations and human-verified evaluation labels.

Method	Spatial R@1 ↑	Spatial mAP ↑
Full SceneBind	**28.9**	**48.0**
No Gaussian smoothing	23.2	43.6
No spatial contrastive	27.0	46.1

Replacing Gaussian-smoothed spatial targets with hard one-hot targets reduces spatial R@1 and mAP, consistent with the predominantly adjacent-bin annotation disagreements in Table 10. Removing cross-batch spatial contrastive supervision also degrades performance. These results support Gaussian-smoothed targets under the observed annotation regime and show a complementary benefit from spatial contrastive learning beyond exact bin assignments.

C.4 Training Stages and Data Recipe

SceneBind is trained in two stages. Stage 1 learns a shared audio-visual-text space from mixed supervision, while Stage 2 specializes the model on Binaural dataset.

Stage 1. Stage 1 trains for 30 epochs with three data sources: Binaural scene dataset, AudioCaps [43], and MS-COCO [44]. Dataset sampling follows

$$p_i \propto n_i^{0.5} w_i,$$

where n_i is the dataset size and w_i is the dataset weight. The paired binaural data receive the largest weight because they provide the full spatially grounded audio-visual signal. AudioCaps provides audio-text semantic supervision, and MS-COCO provides visual-text semantic supervision with partial spatial labels, where only azimuth and elevation supervision are used.

Stage 2. Stage 2 initializes from the Stage 1 model weights, resets the optimizer, and trains for another 30 epochs using only the paired binaural scene dataset.

Optimization and Compute. Both stages use AdamW with weight decay 0.01, cosine learning-rate decay, and gradient clipping at norm 1.0. Stage 1 uses learning rate 10^{-4}, 200 warmup steps, and global batch size 1024. Stage 2 uses learning rate 5×10^{-5}, 100 warmup steps, and global batch size 512. The reported run was trained on one node with 8 NVIDIA H200 GPUs, each with 140GB memory. Using 8 data-loading workers per GPU process, the two-stage run completed in roughly 4 wall-clock hours on this setup.

Table 13: Training data recipe and active supervision. n_i is the source dataset size and w_i is the dataset weight.

Stage	Data source	n_i	w_i	Sample	Active supervision
1	Binaural scenes	38,430	2.0	54.8%	$\mathcal{L}_{\text{global}}$, $\mathcal{L}_{\text{sem}}$, $\mathcal{L}_{\text{conf}}$, $\mathcal{L}_{\theta,\phi,d}$, $\mathcal{L}_{\text{inst}}$, $\mathcal{L}_{\text{batch}}$
1	AudioCaps	91,254	0.5	21.1%	$\mathcal{L}_{\text{global}}$, $\mathcal{L}_{\text{sem}}$, $\mathcal{L}_{\text{conf}}$,
1	MS-COCO	118,248	0.5	24.1%	$\mathcal{L}_{\text{global}}$, $\mathcal{L}_{\text{sem}}$, $\mathcal{L}_{\text{conf}}$, $\mathcal{L}_{\theta,\phi}$
2	Binaural scenes	38,430	1.0	100.0%	$\mathcal{L}_{\text{global}}$, $\mathcal{L}_{\text{sem}}$, $\mathcal{L}_{\text{conf}}$, $\mathcal{L}_{\theta,\phi,d}$, $\mathcal{L}_{\text{inst}}$, $\mathcal{L}_{\text{batch}}$

C.5 Inference Efficiency

We measure end-to-end inference at batch size one on a single NVIDIA H200 using PyTorch 2.9.0 and CUDA 12.8. Measurements use synchronized CUDA timing over 100 distinct clips, with 50 slots, one 256×256 RGB frame, and a two-second binaural input. Peak GPU memory is 4.91 GiB. Encoding a new text prompt takes 6.16 ms; fixed prompts are encoded once and cached. After slot extraction, recurring query and fusion operations take less than 0.23 ms.

Table 14: **Single-GPU inference latency.** Batch size one on an NVIDIA H200, with cached text prompts.

Inference path	Mean latency (ms)	p95 latency (ms)
Audio	6.75	6.81
Visual	7.02	7.12
Audio + visual	13.09	13.45

C.6 SceneBind Matching Details

At inference time, each sample from modality m is represented by $\mathcal{X}^m$, containing a global semantic embedding and a set of object-centric semantic-spatial slots. We first apply a confidence gate, keeping slots with confidence above 0.05 as active slots. If no slot passes the gate, we retain the highest-confidence slot. SceneBind Matching then computes two complementary scores: a global score S_{global} from the cosine similarity between global embeddings, and an object-centric score S_{obj} from matched semantic-spatial slots of paired scenes. We propose different strategy to use these scores for different tasks such as scene retrieval, spatial retrieval and object grounding.

Object Slot Score. Given a query scene q and candidate scene x, we compare each active query slot i with each active candidate slot j using a semantic-spatial score:

$$R_{ij}(q,x) = \cos(\mathbf{s}_i^q, \mathbf{s}_j^x) \cdot \rho_{ij},$$

where $\mathbf{s}_i^q$ and $\mathbf{s}_j^x$ are the semantic embeddings of the two slots. The spatial term is

$$\rho_{ij} = \frac{2\langle \boldsymbol{\pi}_i^{q,\theta}, \boldsymbol{\pi}_j^{x,\theta} \rangle + \langle \boldsymbol{\pi}_i^{q,\phi}, \boldsymbol{\pi}_j^{x,\phi} \rangle + \langle \boldsymbol{\pi}_i^{q,d}, \boldsymbol{\pi}_j^{x,d} \rangle}{4}.$$

Here $\boldsymbol{\pi}^\theta$, $\boldsymbol{\pi}^\phi$, and $\boldsymbol{\pi}^d$ denote the softmax-normalized probability distributions over azimuth, elevation, and distance bins, respectively. The inner products measure distributional agreement along each spatial axis, with azimuth given double weight because horizontal direction is the primary cue in our binaural setting.

Best-Match Object Score. After computing the pairwise slot scores $R_{ij}(q,x)$ between the query scene q and a candidate scene x, each active query slot from q selects the candidate slot from x with the highest confidence-weighted semantic-spatial score:

$$j_i^\star = \arg\max_j w_{ij} R_{ij}(q,x), \qquad w_{ij} = \sqrt{c_i^q c_j^x},$$

where c_i^q and c_j^x are the confidence scores of the query and candidate slots, and w_{ij} is their geometric-mean confidence. We then aggregate the selected matches into an object-centric sample score:

$$S_{\text{obj}}(q,x) = \frac{\sum_i w_{ij_i^\star} R_{ij_i^\star}(q,x)}{\sum_i w_{ij_i^\star} + \epsilon}, \qquad \epsilon = 10^{-8}.$$

Here the summation is over active query slots, and $j_i^\star$ denotes the selected candidate slot for query slot i. Confidence weighting emphasizes reliable slot pairs, while independent best matching supports partial object overlap and variable object counts.

Scene Retrieval. For scene retrieval, we combine the global embedding score with the object-centric score. Let $S_{\text{global}}(q,x) = \cos(\mathbf{s}^q_{\text{global}}, \mathbf{s}^x_{\text{global}})$ denote the global similarity between query scene q and candidate scene x. We first rank candidates by S_{global} and select the top 50 candidates for object-level reranking. For each query, we min-max normalize S_{global} and S_{obj} over the candidate set, denoted as $\bar{S}_{\text{global}}$ and $\bar{S}_{\text{obj}}$. The final scene similarity is

$$S_{\text{scene}}(q,x) = \bar{S}_{\text{global}}(q,x) + \lambda_{\text{obj}} \bar{S}_{\text{obj}}(q,x),$$

for candidates in the top-50 set, while all remaining candidates keep their global score. We use $\lambda_{\text{obj}} = 0.05$ for audio-visual retrieval and $\lambda_{\text{obj}} = 0.5$ for text-related retrieval, since text queries benefit more from object-level semantic-spatial matching.

Spatial Retrieval. Spatial retrieval uses an object clause as the query. The text query provides one-hot spatial distributions $\boldsymbol{\pi}^{t,\theta}$, $\boldsymbol{\pi}^{t,\phi}$, and $\boldsymbol{\pi}^{t,d}$ for azimuth, elevation, and distance. We score a candidate sample x by its best active slot using spatial agreement only:

$$S_{\text{spa}}(t,x) = \max_j c_j^x \cdot \frac{2\langle \boldsymbol{\pi}^{t,\theta}, \boldsymbol{\pi}_j^{x,\theta} \rangle + \langle \boldsymbol{\pi}^{t,\phi}, \boldsymbol{\pi}_j^{x,\phi} \rangle + \langle \boldsymbol{\pi}^{t,d}, \boldsymbol{\pi}_j^{x,d} \rangle}{4},$$

where c_j^x is the confidence of candidate slot j, and $\boldsymbol{\pi}_j^{x,\theta}$, $\boldsymbol{\pi}_j^{x,\phi}$, and $\boldsymbol{\pi}_j^{x,d}$ are the softmax-normalized predicted distributions over azimuth, elevation, and distance bins.

Object Grounding. For object grounding, the query is an object-clause text semantic description encoded as a text embedding $\mathbf{s}^t$. We first apply the same confidence gate to the predicted slots of the queried scene, retaining slots with $c_i > 0.05$ and falling back to the highest-confidence slot if none pass. Among the retained slots, we select the slot with the largest semantic-confidence score:

$$i^\star = \arg\max_i c_i \cos(\mathbf{s}_i, \mathbf{s}^t).$$

If the selected slot has confidence below 0.05 or semantic similarity below 0.5, the queried object is treated as missing. Otherwise, we take the argmax of the selected slot's azimuth, elevation, and distance logits as the predicted spatial bins, and compare them with the ground-truth labels.

Table 15: Fine-tuning losses for global-embedding baselines. Global losses use one paired embedding per sample, while clause losses use all valid object-clause embeddings associated with the sample.

Loss term	Supervision source	Weight
$\mathcal{L}^{a,v}_{\text{global}}$	Paired audio and visual sample embeddings	2.0
$\mathcal{L}^{a,t}_{\text{global}}$	Audio embedding and scene-description text embedding	1.0
$\mathcal{L}^{v,t}_{\text{global}}$	Visual embedding and scene-description text embedding	1.0
$\mathcal{L}^{a,t}_{\text{clause}}$	Audio embedding and object-clause text embeddings	0.5
$\mathcal{L}^{v,t}_{\text{clause}}$	Visual embedding and object-clause text embeddings	0.5

D Baseline Finetune Details

We fine-tune all retrieval baselines on the same data sources and splits as SceneBind, following Table 13. This gives all methods access to the same audio-visual samples, scene descriptions, and object-clause annotations. Unlike SceneBind, the baselines produce a single embedding per modality sample and have no spatial grounding heads, so spatial attributes cannot be supervised as separate structured labels. We therefore incorporate the spatial attributes into the object-clause text itself, yielding textual queries that describe both object semantics and spatial location. This setting is used for all baselines in Sec. 4.3.

Zero-shot Alignment. CLAP-based audio encoders and SigLIP2 align audio-text and visual-text separately, making their zero-shot audio-visual alignment indirect. AudioCLIP was trained for broad category-level alignment, while ImageBind directly aligns audio and vision during pretraining. All baselines use their native audio preprocessing. The improvement under matched fine-tuning reflects the pretraining-domain and cross-modal alignment gaps.

Training Objective. For paired modalities m and n, each baseline is optimized with the same symmetric InfoNCE objective used for global alignment. We apply this loss to audio-visual, audio-text, and visual-text pairs whenever the corresponding supervision is available. Since baselines produce one embedding per modality sample rather than object slots, we use object-clause annotations through a multi-positive clip-to-clause contrastive loss.

Let $\mathbf{s}_b$ be the normalized embedding of sample b in a minibatch of size B, and let $\{\mathbf{t}_\ell\}_{\ell=1}^{L}$ be all normalized object-clause text embeddings in the minibatch. We denote by $o(\ell)$ the sample that owns clause ℓ. The multi-positive clause loss is

$$\mathcal{L}_{\text{clause}} = \frac{1}{2}\left[-\frac{1}{B}\sum_{b=1}^{B}\log\frac{\sum_{\ell:o(\ell)=b}\exp(\mathbf{s}_b^\top\mathbf{t}_\ell/\tau)}{\sum_{\ell=1}^{L}\exp(\mathbf{s}_b^\top\mathbf{t}_\ell/\tau)} + \text{CE}\left(\frac{TS^\top}{\tau}, o\right)\right], \qquad \tau = 0.07.$$

Here S stacks the B sample embeddings, T stacks the L clause embeddings, and o is the clause-to-sample target index. The first term pulls each sample toward all of its own object clauses, while the second term assigns each clause back to its owning sample.

The clause loss treats all object clauses from the same sample as positives for the corresponding sample embedding, and also assigns each clause back to its owning sample. Table 15 summarizes the baseline fine-tuning losses and their weights. Thus, baselines receive both global-level and object-clause text supervision, but do not perform object-centric prediction or spatial decoding.

ImageBind. For ImageBind [8], we use the pretrained ImageBind audio and image encoders to obtain audio and visual embeddings. We add lightweight trainable projection heads to map these embeddings into the shared retrieval space. The pretrained backbone is kept frozen in the standard fine-tuning setting, and only the retrieval heads are optimized.

M2D+SigLIP2. For M2D [7]+SigLIP2 [5], audio is encoded with the pretrained M2D-CLAP audio encoder and visual/text features are encoded with SigLIP2. A trainable retrieval head maps the audio representation into the SigLIP2-aligned retrieval space. In the fair comparison setting, the pretrained M2D backbone and the audio-to-SigLIP2 alignment module are frozen, and only the retrieval head is trained.

SpatialCLAP+SigLIP2. For SpatialCLAP [10]+SigLIP2 [5], binaural audio is encoded by the pretrained SpatialCLAP audio encoder, while visual and text features are encoded by SigLIP2. We

add trainable projection heads to align the audio and visual embeddings in a common retrieval space. The pretrained audio alignment module is frozen in the fair comparison setting.

All baselines follow the same two-stage fine-tuning schedule as SceneBind. Stage 1 trains for 30 epochs on the mixed data recipe with paired Binaural dataset, AudioCaps, and MS-COCO, with a batch size of 1024. Stage 2 initializes from the Stage 1 weights and trains for another 30 epochs on the paired Binaural dataset only, with batch size equals to 512. We use AdamW with weight decay 0.01, cosine learning-rate decay, warmup, and gradient clipping at norm 1.0. The learning rate is 10^{-4} in Stage 1 and 5×10^{-5} in Stage 2, matching the SceneBind schedule (Table 13 and Sec. C.4).

Baselines are evaluated with the same retrieval pools, benchmark splits, and sample-level protocol as SceneBind. At inference time, each baseline produces a single embedding per modality sample. To keep the scoring protocol consistent with SceneBind Matching C.6, we use this embedding both as the global representation and as a one-slot object representation for object-level matching. Since the baselines do not predict spatial attributes, they cannot explicitly compare azimuth, elevation, or distance labels. For spatially conditioned text clause queries, we instead encode the combined semantic-spatial description as one text embedding and match it directly to the sample embedding.

E Balanced Spatial Evaluation and Binaural Cue Ablation

Evaluation metrics. We complement per-attribute micro accuracy with macro recall and majority-class rates to account for spatial class imbalance. Micro accuracy is computed over successfully grounded objects. Macro recall averages recall across occupied ground-truth bins, counting unmatched objects as false negatives; empty bins are excluded. Majority-class rates indicate the prevalence of the most frequent spatial category.

Controlled mono comparison. To isolate interaural cues, we average the left and right waveforms and duplicate the mono signal across both spatial-encoder inputs during training and evaluation. We keep the backbone, architecture, parameter count, training schedule, and random seed fixed.

Table 16: **Audio spatial evaluation under binaural and mono input.** Single-seed controlled comparison. Micro and macro denote accuracy over successfully grounded objects and recall averaged across occupied ground-truth bins, respectively.

Attribute	Binaural micro	Mono micro	Majority class	Binaural macro	Mono macro
Azimuth	39.4	40.6	41.7	19.9	16.8
L/R	62.7	50.8	52.2	59.5	46.2
Elevation	83.2	83.9	83.8	36.8	30.1
U/D	77.8	76.2	53.8	73.0	69.6
Distance	58.4	59.6	37.7	49.8	50.3

The strongest binaural benefit is in horizontal discrimination: L/R macro recall increases from 46.2% to 59.5%. Elevation micro accuracy is dominated by the frequent level bin, whereas macro recall increases from 30.1% to 36.8%. Distance shows no corresponding binaural gain, consistent with its reliance on reverberation, attenuation, and contextual cues that remain available in mono audio.

Per-bin elevation analysis. Binaural input improves recall in every non-level elevation bin while preserving similar level-bin recall (Table 17). These results show that elevation prediction benefits from binaural cues despite the strong class imbalance.

Table 17: **Audio elevation recall by bin.** E2 denotes the majority level bin.

Elevation bin	Binaural recall	Mono recall
E0	34.0	27.7
E1	10.8	1.4
E2 — level	90.9	91.3
E3	28.1	23.6
E4	20.0	6.7

F Zero-Shot Audio-Visual Localization Details

For the zero-shot audio-visual localization task in Sec. 4.7, the input is a stereo audio clip and an image frame. We evaluate SceneBind without finetuning. The audio branch is run once on the stereo audio to produce audio object slots, while the visual branch is run on the image to produce visual object slots and the global semantic query's cross-attention over visual patch tokens. This global cross-attention map provides a coarse visual support map for candidate sounding-object regions.

In this task, the sounding object should match a visual object in semantics. We therefore score audio-visual slot pairs by semantic similarity and slot confidence, and use the best-matched audio slot to weight visual cross-attention maps. We then construct spatial guide maps for the selected pair on the image plane. Since camera intrinsics are unavailable, we use a normalized camera-coordinate projection: the image center corresponds to azimuth 0° and elevation 0°, horizontal position is determined by azimuth, and vertical position is determined by elevation. Under this projection, the selected visual slot's predicted azimuth and elevation define a center on a 72×72 grid, where we render a 2D Gaussian guide map with $\sigma_x = \sigma_y = 0.13$.

For the selected audio slot, we build an analogous guide map from its predicted spatial distributions. We temperature-scale the azimuth, elevation, and distance logits with various temperatures (e.g., 0.45, 0.60, and 0.75), respectively, before softmax normalization. The audio guide center is estimated from the predicted azimuth/elevation bin and its immediate neighboring bins. The distance distribution controls the Gaussian width through distance-dependent base scales, with a small confidence-dependent broadening.

The visual and audio guide maps are normalized and fused into an audio-visual support map, indicating where the matched audio-visual slot pair agrees the sounding object should lie in the image. We use this support map to select the relevant region from the visual cross-attention map. The cross-attention map is thresholded at multiple levels to produce connected-component candidates, and each component is scored by its overlap with the support map and the strongest audio-visual overlap seed. We keep the highest-scoring component and suppress the rest. The final localization map therefore preserves the structure of the visual attention map while selecting the region most consistent with the matched audio-visual spatial evidence.

For cIoU evaluation, the final heatmap is normalized and thresholded at 0.4. If multiple connected components remain, we keep the component with the strongest support from the audio-visual guide map. The resulting binary mask is evaluated against the ground-truth bounding-box mask.

G Discussion

Semantic-spatial representation. SceneBind learns a joint sense of what sources are present and where they lie in camera-relative space. Its object-centric slots combine semantics, spatial distributions, and confidence, providing a lightweight spatial prior for downstream tasks.

Learning from weak spatial supervision. The training annotations provide coarse cross-modal supervision. Gaussian-smoothed targets preserve adjacency between spatial bins, while intra- and cross-scene contrastive objectives align compatible objects and separate different instances. Evaluation on human-verified annotations tests generalization beyond the automatic labeling pipeline.

Global and object-level structure. Global embeddings capture holistic scene context, while slots preserve source-level semantics and spatial layout. Structured matching exposes this learned object-level information for retrieval and grounding.

H Limitations and Future Works

Data Scale and Annotation Quality. SceneBind relies on curated spatially aligned audio–visual–text data with structured semantic and spatial annotations. While our binaural dataset provides a strong starting point, its scale and diversity remain limited compared to web-scale semantic datasets. Real-world scenes exhibit long-tail event distributions, complex acoustic conditions, and ambiguous spatial cues that are difficult to capture with current annotation pipelines. In addition, our automatic annotation step is guided by Gemini, which does not directly perceive binaural spatial cues. Its

spatial estimates are inferred mainly from visual evidence, which can bias annotations toward visible or front-facing objects and may miss or hallucinate off-screen or rear-field events. Although we mitigate these issues with cross-modal verification, data balancing, and human review for evaluation, improving automatic spatial annotation quality remains an important direction. Future work will scale data collection with richer in-the-wild sources, strengthen multimodal filtering and human-in-the-loop refinement, and explore semi-supervised or self-supervised learning to reduce reliance on explicit spatial labels.

Spatial Resolution. SceneBind learns coarse camera-relative spatial categories. These categories provide useful spatial supervision and a lightweight prior for downstream localization. Precise metric localization requires calibrated sensors and datasets with accurate depth, camera geometry, source positions, and object tracking. Adapting SceneBind to such datasets is an important direction for higher-resolution spatial modeling. Sphere360 provides known listener rotations. Validation with measured source positions remains an important direction.

Spatial Audio Rendering. SceneBind is trained on screened in-the-wild binaural recordings. Sphere360 is used only for zero-shot evaluation, with listener orientations rendered from real ambisonic recordings. Its fixed KEMAR HRTF does not capture the full range of listener and device variability.

Temporal Modeling. The current formulation operates on short temporal windows (e.g., a single visual frame and $\sim$2s audio), which limits the ability to capture temporal dynamics such as object motion, evolving interactions, and long-range semantic-spatial representation. Extending SceneBind to incorporate temporal integration over longer and flexible time horizons is an important direction. This includes modeling continuous view or object trajectories, temporally consistent slot representations, and cross-time semantic–spatial reasoning for dynamic scenes.

I Broader Impacts

SceneBind studies in-the-wild semantic-spatial omni-modal representation learning, aiming to move multimodal perception beyond recognizing what is present toward understanding where objects and events are located in 3D space. Such omnidirectional spatial perception can support more natural spatial reasoning in assistive systems, wearable devices, AR/VR interaction, multi-modal navigation, human-robotic interaction, and embodied agents. The representation can also serve as a structured interface for larger systems: combined with language models, it may support natural spatial question answering and instruction following; combined with generative or world models, it may help construct semantically and spatially aligned scene representations, or provide a reward or critic signal for evaluating multimodal spatial consistency and improving omni-modal generative models.

At the same time, models that infer spatial structure from audio-visual input can create risks if deployed without safeguards. Potential misuse includes surveillance, unauthorized tracking, or context inference about people and environments without consent. SceneBind also inherits limitations and biases from its pretrained encoders and training data, and its spatial predictions may be incorrect in ambiguous or safety-critical settings. To reduce these risks, deployments should communicate uncertainty and failure modes, follow privacy-preserving data collection and evaluation practices, and restrict use in sensitive applications involving biometric data, private spaces, or real-time monitoring. Future work should further study robustness, interpretability, and consent-aware evaluation protocols for spatial omni-modal perception and reasoning.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims precisely reflect the paper's contribution and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mentioned the limitations in the Conclusion section and appendix.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: Our work is experiment-driven.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We incorporated sufficient details of our pipeline, baseline methods, evaluation metrics and etc, in both the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We commit to releasing the implementation code, evaluation scripts, prompts, annotations, timestamps, source IDs, and split manifests, subject to licensing and takedown requirements. Raw media will not be redistributed without an appropriate license or explicit permission.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specified all implementation details of our proposed pipeline and baseline methods as well as evaluation metrics necessary to achieve the results we claim in the experiments, in both the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Sec. 4.2 reports three-seed means and sample standard deviations for SceneBind's Binaural scene retrieval, spatial retrieval, object grounding, and Sphere360 retrieval. The standard deviations measure variation across seed runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: Training hardware and runtime are specified in Appendix Sec. C.4; inference hardware, input sizes, software versions, latency, and peak memory are reported in Appendix Sec. C.5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and carefully followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed broader impacts in our appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We discussed the need for safeguards regarding the negative impact associated with our method and propose precautions and guidelines to prevent misuse. These details are included as part of the Broader Impact section in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited all open-source works that we use including the data, models and codes.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We clearly described our new benchmark and dataset in the paper, including the details of statistics and distributions, how we curate the dataset in details.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We included detailed description of human study instructions (specifically the study is for benchmark verification), and and the setup of the of the study in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: There are no risks associated with our human verification process.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We explained how we used LLMs (Gemini) to curate the dataset. The details can be found in our appendix.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.